This is to certify that the

dissertation entitled

The Effects of Rating Format and Rater Training on Performance Rating Accuracy and the Motivation to Rate Accurately.

presented by

Robert Lloyd Heneman

has been accepted towards fulfillment of the requirements for

Ph.D. degree in Labor and Industrial Relations

Michael L. Moore
Major professor
Michael L. Moore

Date February 15, 1984

THE EFFECTS OF RATING FORMAT AND RATER TRAINING ON PERFORMANCE RATING ACCURACY AND THE MOTIVATION TO RATE ACCURATELY

By

Robert Lloyd Heneman

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

School of Labor and Industrial Relations

1984

ABSTRACT

THE EFFECTS OF RATING FORMAT AND RATER TRAINING ON PERFORMANCE RATING ACCURACY AND THE MOTIVATION TO RATE ACCURATELY

By

Robert Lloyd Heneman

An important criterion in the evaluation of a performance appraisal system is the accuracy of performance ratings. Two methods of increasing rate accuracy, rating format and rater training, and their impact on the motivation to rate accurately, were considered in this dissertation.

It was hypothesized, based upon cognitive processing theory and expectancy theory, that performance rating accuracy and the motivation to rate accurately would be greater when: (1) behaviors rather than traits were rated; (2) observational rather than rater error training was provided; and (3) the rating format and rater training were consistent with another. Finally, it was expected that there would be a positive correlation between performance rating accuracy and the motivation to rate accurately.

These hypotheses were tested in a laboratory experiment with 87 supervisors from a western utility company. A 2 x 3 factorial design was used. The first factor, rating format, consisted of two levels: behavior scale and trait scale. The second factor, rater training, was defined by three levels: rater error training, observational training, and

control group training. The results of this experiment provided no support for the hypotheses. Instead, it was found that traits were rated more accurately than behaviors.

Two conclusions were made on the basis of these data. First, it appears that raters cognitively process performance information using trait oriented schema. Consequently, their ratings are more likely to be accurate when traits rather than behaviors are rated. Second, these findings suggest that raters are highly motivated to make accurate ratings. Therefore, greater emphasis should be placed upon increasing the skill levels rather than the motivational levels of the rater. Both sets of conclusions must, however, be treated as tentative given the methodological limitations associated with this study.

ACKNOWLEDGMENTS

This is the highlight of my experiences as a graduate student. It is an opportunity to formally thank those individuals and institutions that have helped me to complete this dissertation, to achieve my academic goals, and to grow as a person. While all of the following parties have provided me with assistance in all three of these endeavors, each party has had a unique impact that I would like to acknowledge and thank.

I owe my initial interest and guidance in the field of industrial relations to Milton Derber, Herbert Heneman, Jr., and Kendrith Rowland.

The faculties and students at the Institute of Labor and Industrial Relations, University of Illinois, and the School of Labor and Industrial Relations, Michigan State University, helped me to develop an interdisciplinary view of the employment relationship.

Several individuals have shared keen insights with me into various subelements of the field: Richard Block, Thomas Patten, Michael Moore, Neal Schmitt, and Kenneth Wexley. In addition, they made many helpful comments and suggestions as members of my dissertation committee.

The University of Illinois at Urbana-Champaign and Michigan State University have provided me with the

and resources that I have needed as a graduate student. Funding provided by the American Compensation Association and the western utility company used in this experiment made the completion of this dissertation possible.

I have received a great deal of administrative assistance from Richard Block, Einar Hardin, and Michael Moore in the preparation of my program and the dissertation. I received technical assistance in various stages of this project from Leslie Corbitt, Rob MacCoun, Paul Reagan, and Bob Taylor. In addition to providing technical assistance, Steven Premack made many insightful comments on an earlier draft of this paper.

Michael Moore, Bob Taylor, and Kenneth Wexley have picked me up when I needed it, kicked me in the pants when it was necessary, and patiently provided me with guidance and counsel through the entire project.

Finally, I have been blessed by the strength and encouragement given to me by my Mother and Father; my brother, sister, and their families; Richard and Marihelen Davis; Mado Kreutz; and a number of friends in California, Illinois, Michigan, and Minnesota.

TABLE OF CONTENTS

Page

Page

APPENDICES (Cont.)

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction and Overview

This chapter provides a brief introduction to and overview of the entire dissertation. This will be accomplished by stating the research objectives, looking at the reasons why research on this topic is important, and briefly describing the content of each of the six chapters to follow.

Research Objectives

Many organizations rely upon performance ratings for a number of personnel decisions including pay, promotions, and layoffs (Bureau of National Affairs, 1983). In order for a performance appraisal system to be successful it is of major importance for raters to have the skills and motivation necessary to make accurate ratings (Bernardin & Cardy, 1982). Two methods of increasing rating accuracy, rating format and rater training, are considered in this dissertation. Unlike previous research in this area, the following propositions are set forth and are then tested:

1. Performance rating accuracy is a function of the ability and the motivation of the rater. Very little attention has been given to the motivation component. In this study the relationship between rating accuracy and the motivation to rate accurately is assessed.

2. Rating format and rater training affect performance rating accuracy and motivation to rate accurately. Previous research has ignored the impact of rater training and to a lesser extent, rating format, on the motivation to rate accurately. Consequently, the effects of these two independent variables, on the motivation to rate accurately, are examined. Expectancy theory is used to explain why this effect is to be expected.

3. Rater training and rating format cannot be considered independent of one another as has been the case with previous research. The interactive effect of these two variables may account for a significant amount of variance in rating accuracy and motivation to rate accurately. Cognitive processing theory and expectancy theory are used to explain this hypothesized, interactive effect.

4. Given the limited understanding of the cognitive processing of performance information by raters, there has been too much emphasis placed upon minimizing common rater errors like halo and leniency, and not enough emphasis placed upon developing the observational skills of raters (e.g. gathering critical incidents). A new type of training, observational training, is set forth here and is tested against rater error training and a comparison group receiving a general overview of performance appraisal.

Importance of the Topic

The objectives of this dissertation are of importance to both performance appraisal researchers and practitioners.

For the former group, this dissertation addresses several recent calls in the literature for future research:

1. Several authors have indicated that a cognitive processing view is needed to better understand the rating process (e.g. Atkin & Conlon, 1978; Borman, 1978; Feldman, 1981; Landy & Farr, 1980; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982). To date there have been very few studies taking this approach. This study is grounded in cognitive processing theory.

2. Researchers have been criticized for taking too narrow a view of rating accuracy by concentrating on either rating format or rater training (Zedeck & Cascio, 1982). A broader view is taken here by looking at the interaction between these two independent variables; and the end result is a step towards a contingency view of performance appraisal (Keeley, 1978). The type of rater training to be used can be matched with the type of rating format being used.

3. While several authors have called for rater training programs emphasizing observational skills (Bernardin & Buckley, 1981; Borman, 1979 a; Landy & Farr, 1980; Spool, 1978), none have been developed or at least have not been reported in the published literature. This omission is particularly noteworthy as observation is an important cognitive task confronting the rater (Feldman, 1981). An observational training program was developed for this dissertation based upon the theory of observation set forth by Weick (1968). Moreover, this program was tested against

rater error training and a comparison group receiving a general overview of the performance appraisal process.

4. DeCotiis and Petit (1978) and Mohrman and Lawler (1983) have emphasized the importance and determinants of motivation to rate accurately. Little research has been conducted along these lines. This study examines the impact of rating format and rater training method on motivation to rate accurately. In turn, the relationship between motivation and rating accuracy is assessed.

This dissertation also addresses the concerns of practitioners as indicated in the following points:

1. Downs and Moscinski (1979) surveyed 67 directors of training in Fortune 250 companies and found that these respondents were very concerned by the fact that there was "subjectivity in the ratings" for those raters using their present appraisal system and that the "appraiser's skills are underdeveloped." These two conclusions emphasize the need for further research on performance rating accuracy and training.

2. Baird (1982) provides three reasons why performance rating accuracy should be of importance to practitioners. First, he feels that they are essential to the management of performance. Without an accurate criterion measure, it is impossible to ascertain whether goals have been achieved and very difficult to provide performance or development counseling. Second, he points to the Uniform Guidelines and several court cases which emphasize the importance of

accurate ratings. Finally, many human resource management subsystems are dependent upon accurate performance ratings. For instance, it is very difficult for organizations to make the link between employee performance and rewards without accurate ratings (Lawler, 1971).

3. Suspicion concerning the accuracy of ratings may be a stumbling block to getting supervisors to even use rating instruments (McGregor, 1957). It might be possible to begin to overcome this obstacle by providing supervisors with rating formats and training programs that will provide them with the ability and motivation to make accurate ratings.

Chapters in the Dissertation

Presented below is a brief description of the material contained in the next six chapters of this dissertation.

Literature Review

This chapter presents a review of the literature on the accuracy of performance ratings. The following topic areas are covered: definitions of accuracy, models of accuracy, empirical studies of accuracy that have been conducted, and methods used to develop 'true' scores for the calculation of accuracy. Within each of these areas emphasis is placed upon conceptual and methodological issues, and on future research directions.

Models and Hypotheses

The major hypotheses tested in the dissertation and the models they were deduced from are presented in this chapter.

There are three major sections. The first one is concerned with the accuracy of performance ratings. A brief review of cognitive processing models is presented and from these models, several hypotheses are developed concerning the impact of rating format and rater training on performance rating accuracy. The second section deals with the motivation to rate accurately. An expectancy model of motivation is discussed as it relates to performance ratings and hypotheses are generated about the anticipated effects of rating format and rater training on the motivation to rate accurately. In the final section, the expected relationship between performance rating accuracy and the motivation to rate accurately is presented.

Research Methodology

This chapter describes the experimental design used, subjects recruited, procedures undertaken to test the hypotheses, and the methods of data analysis. A detailed description of the rating formats, rater training programs, and dependent measures is provided.

Results

The analysis of variance, effect size, reliability, and correlational results are reported in this chapter. A description of the support or lack of support for each hypothesis is presented.

Discussion

In this final chapter a discussion is presented concerning the support or lack of support for the hypotheses, the theoretical and applied implications of this research, the limitations associated with this study, and the directions that future research in this area might take.

CHAPTER 2

Literature Review

Given the importance of accurate performance ratings to human resource managers and to students of performance appraisal, a surprisingly small amount of theory and evidence has been generated on this topic. In this chapter, the available literature will be reviewed by looking at the definitions of accuracy that have been offered, models and theories that have been set forth, empirical studies that have been conducted, and methods used to develop 'true' scores for the calculation of accuracy.[1] Each of these topics will be covered in turn. At the end of each section the implications will be discussed. These conclusions serve as the stimulus to the theory and hypotheses developed in the next chapter.

Accuracy Definitions

In a very general sense, performance rating accuracy has to do with the relationship between actual employee

---

[1]It should be noted at the outset that an attempt was made to be comprehensive at reviewing those studies concerned with the accuracy of performance ratings. Given the scope of this project and questions concerning the generalizability of findings, no such attempt was made to review all of those studies concerned with person-perceptions in general or eyewitness accuracy.

behavior and employee behavior that has been recorded by a rater. Gordon (1970) offers a more precise definition:

> Accuracy is a function of the total amount of error inherent in an instrument. This includes both variable error which is measured by an index of dispersion and constant error, which is a function of the difference in the location of the distributions obtained with the fallible (performance rating) and less fallible (actual employee behavior) instruments ..., p. 367.

Hence, accuracy is made up of two error components: random and constant error. When these two sources of error variance are minimized, a measure is said to be accurate.

Given this conceptualization of rating accuracy, two sets of measures have been set forth to operationalize this construct. The first set of measures are concerned with variable error. More specifically, the correlation between actual employee behavior (true score) and the performance rating of employee behavior (observed score) is calculated. The second major approach is based upon constant error. Here, the distance between the true and observed score is calculated. A narrative description of specific measures within each of these two categories and uses for each measure will now be presented.[2]

Correlational Measures

Cronbach (1955) set forth several definitions of accuracy which focus on variable processes. The first

---

[2]Readers interested in the development and use of statistical formulas for each measure are referred to Borman (1979a), Cronbach (1955), Murphy, et. al. (1982), and Wiggins (1973).

definition, differential evaluation, gives a measure of association between the ordering of each employee by their true performance and the ordering of employees made by the rater. This measure of accuracy is important when the rater is required to identify the best performers in his work group (Murphy et. al., 1982). He may need to do so for a variety of personnel decisions including merit pay and promotion.

The other two definitions developed by Cronbach (1955), stereotype accuracy and differential accuracy, are somewhat similar. Both are concerned with degree to which the rater's judgements covary with the true performance profile of the employee(s). Stereotype accuracy is important when, for example, the rater must assess the skill deficiencies of his employees in order to select a training program (Murphy, et. al. 1982). Hence, this definition is concerned with the performance profile of the group. On the other hand, differential accuracy focuses on performance profiles for each ratee. As a result, it is an important consideration when the rater is charged with making placement or job assignment decisions (Murphy et. al., 1982). In these situations the rater must match the performance of the ratee along a number of dimensions with the performance dimensions required by the job.

Difference Score Measures

There is one major definition of accuracy, elevation (Cronbach, 1955), that takes into explicit consideration the

distance between observed and true scores. As originally conceived (Cronbach, 1955), accuracy was defined as the distance between the rater's average score for a group of ratees and the true score average. This definition is of importance when the rater is asked to make distinctions between the performance of work groups within his control (Murphy et. al. 1982). For instance, the vice president of human resources may be asked to allocate rewards to various subunits within personnel (e.g. recruitment, compensation, etc.) on the basis of subunit performance.

Other variations of this distance notion also exist. One common variation is a 'hits' or 'misses' definition where accuracy is defined as the number of correct or incorrect rating judgments made by the rater. This measure may be appropriate, for example, in a discipline situation where the rater is asked to make a judgment concerning the number of times a certain rule infraction occurred.

While the number of hits or misses may be important in some situations like discipline, there is a greater concern with how close the rater's observation is to the true score (Naylor, 1967). Hence, some authors (e.g. Heneman & Wexley, 1983) have used an elevation score where the distance between the rater's score and the true score on each item for each ratee is assessed. This approach would appear to be important when the rater is concerned with performance ratings for employee development purposes. That is, the rater is providing feedback and guidance to each ratee

concerning their progress toward a number of predefined goals or behavioral standards.

Accuracy Measures Contrasted

According to Borman (1977), the question as to which definition of rating accuracy is most appropriate is a closed one as demonstrated by this quote: "Differential accuracy definitely appears to be the most appropriate for assessing the accuracy of performance judgements, p. 240." It will be argued here that when one looks at the purpose of the rating, tradeoffs involved in focusing on definitions emphasizing variable or constant error, and various statistical considerations, there is no one best definition.

Examples given for each of the definitions depends upon the purpose of the appraisal. Differential elevation is needed when employees must be rank ordered for personnel decisions. Stereotype accuracy appears to be the proper definition when the training needs of a work group are to be assessed. If the rater is required to make placements or job assignments, the differential accuracy definition is more suitable. When reward allocations are to be made to various groups, the elevation definition offered by Cronbach (1955) is needed. A hits or misses definition can be used when the rater is asked to specify the number of times a certain behavior occurred. Finally, an absolute difference score for each item is appropriate for assessing the accuracy of employee development needs and progress.

Not only does the purpose of the rating determine the appropriateness of the accuracy measure, but so do considerations concerning variable and constant error (Cronbach, 1955). Using correlational definitions, accuracy is the degree to which true and observed scores covary with one another. While this covariation is important in some situations, it is not always. Even if true and observed scores correlate perfectly with one another, they may be a great distance apart from one another (Tinsley & Weiss, 1975). Thus, for instance, the rater may have an excellent view of the pattern of employee behaviors, but may greatly over-or under-estimate the behaviors.

Turning to difference score measures, the emphasis is on how close the observed scores are to the true scores. As a result, the rater's observed scores may, overall, be close to the true scores, but distort the actual pattern of behaviors. Which approach is best depends upon the purposes of the rating. Covariation may be important in some situations while distance may be important in others.

Finally, several statistical issues must be taken into consideration. First, reliability problems with difference scores are well documented (e.g. Cronbach & Furby, 1970). While this problem may not be as severe as initially thought (Rogasa, Brandt, & Zimowski, 1982), it does give an edge to correlation rather than difference score measures of accuracy. Second, Cronbach's (1955) definitions assume that there are multiple ratees (Richards & Cline, 1963). When

this is not the case (i.e. in research where only one ratee is rated) the original formulas are no longer applicable. Third, at least two studies (Murphy et. al., 1982, & Richard & Cline, 1963) have found that some of the measures of these definitions do correlate with one another. Hence, at a practical level, it may be possible to substitute one measure for another.

In summary, correlational and distance definitions of performance rating accuracy have been advanced. Within each approach a number of different measures are possible depending on whether the ratings are averaged across raters, ratees, and dimensions. The choice of a definition and measure depends upon the purpose of the measurement and the rating, and on several methodological considerations. Future research in this area might be directed toward a better understanding of the interrelationships of the measures within and between these two approaches. Moreover, those developing training programs to increase rating accuracy might want to look to these definitions to determine the content of the program. Training, for example, to increase accuracy defined by a correlational measure may need to be different than training to increase a distance measure of accuracy.

Accuracy Models

Four models have been set forth that treat accuracy as the dependent variable. These models will be briefly reviewed and then a general discussion of the strengths and

shortcomings of each model will be presented at the end of this section.

Spool (1978)

Based on the work of Cronbach, Gleser, Nada, and Rajaratnam (1972), a model of accuracy was presented by Spool (1978). Accuracy was depicted as a function of three factors: Recording procedure characteristics, observer characteristics, and conditions of observation. Components of the recording procedure include the format used, the complexity of the format, and how ratings are recorded. Observer characteristics include the age, sex, expectancies, intelligence, and rating experience of the rater. Conditions of observation include the characteristics of the ratee, the number of ratees, the behaviors that occur, the frequency and rate at which behaviors occur, and the temporal sequencing of behaviors. No attempt was made to predict the strength and the direction of the relationship between these variables and rating accuracy.

Ilgen (1983)

A more specific model of rating accuracy was presented by Ilgen (1983). It was hypothesized that overall rating accuracy is a function of the objectivity of performance standards, the appraiser's knowledge of the dimensions to be rated, the opportunity to observe, and the expectations of the rater for employee performance. These variables are very consistent with those set forth by Spool. These

factors are thought to have a direct influence on overall accuracy. Variables which have an indirect influence on overall accuracy, primarily through their impact on the variables just listed, include attribution sex effects, past experiences, and sex-role expectations.

Ilgen then goes on to describe variables that influence two special errors in rating accuracy: over - and under - estimation. The underestimation of performance is directly influenced by appraiser expectations for appraisee performance and indirectly influenced by past experience. Performance that is overestimated was hypothesized to be directly effected by appraiser/appraisee similarity and appraiser expectations for appraisee performance. The overestimation of performance was thought to be indirectly effected by past experiences and sex-role effects. In most cases Ilgen specified the strength and direction of these relationships.

DeCotiis and Petit (1978)

In very general terms accuracy was conceived of as a function of the rater's ability and motivation, and the availability of appropriate rating standards. The rater's ability or "...skill with which a rater interprets job behavior, p. 639" is dependent upon the opportunity to observe ratee behavior, characteristics of the rater, training received by the rater, and the availability of appropriate rating standards. Motivation or what energizes, directs, and sustains energy for accurate ratings is deter-

mined by the perceived consequences of an accurate appraisal for the rater and ratee, the adequacy of the rating format as perceived by the rater, the rating format, the purpose of the appraisal, and the availability of appropriate performance standards. Finally, the availability of appropriate rating standards was hypothesized not to have a direct effect on accuracy. Instead the authors felt that it indirectly effected accuracy through the motivation and ability of the rater. The availability of appropriate rating standards is a function of the job characteristics and the personality of the ratee, the appraisal format, and the organizational policies and procedures for performance appraisal. The direction of each of these relationships were predicted by the authors.

Wherry and Bartlett (1982)

This psychometric theory of rating accuracy was initially developed by Wherry (1952) and then edited and commented on by Bartlett. It was assumed that the accuracy of ratings is a function of the following processes: performance by the ratee, observation of the performance by the rater, and recall of the observations by the rater. From these assumptions, a theory of rating accuracy was developed and can be summarized with the following equation (Bartlett, 1983):

$$z_R = W_T z_T + W_B z_B + W_I z_I + W_E z_E \quad (1)$$

where:

z_R = rating accuracy

$W_T Z_T$ = true ability of the rater

$W_B Z_B$ = bias of the rater

$W_I Z_I$ = environmental influences

$W_E Z_E$ = random error

Accurate ratings occur when the weight given to the true ability variable is maximized and when the weights given to the other variables are minimized.

From a decomposed version of this equation, a total of 17 theorems and 23 corollaries were deduced. Major variables identified in the theorems, which have an impact upon the weights of the variables in the decomposed equation, included the following:

- control over the task by the ratee
- observability of rating scale items
- training concerning what activities are to be rated
- conscious effort to be objective
- checklist of objective cues for the evaluation of performance
- physical features of the scale that facilitate recall
- diary of critical incidents
- importance of the rating to the ratee and society
- knowledge that the rating will have to be justified
- delay time between observation and recall
- intention to remember
- performance and rating items that are easily classified into categories

- number and relevancy of previous contacts with the ratee

The strength and direction of the relationship between these variables and rating accuracy were fully specified by the authors.

Accuracy Models Contrasted

A comparison of these four models reveals a number of commonalities, a number of strengths and weaknesses for each model, and a more comprehensive view of those variables impacting rating accuracy.

In all four models the dependent variable, accuracy, is never fully defined. As shown in the definitions section of this chapter there are a variety of definitions with very different implications. Predictions are difficult to make using these models because it is difficult to know whether the authors are trying to predict correlational or difference score accuracy. The antecedents of these definitions may not be the same. Ilgen (1983) comes the closest to offering a precise definition of rating accuracy in his development of the determinants of under- and over-estimates of performance. This implies a difference score definition of accuracy.

Another common theme to all four models is the emphasis placed upon the ability of the rater. Ability is defined with factors like the intelligence of the rater (Spool, 1978), appraiser's knowledge of the performance dimensions (Ilgen, 1983), and training received by the rater (DeCotiis

and Petit, 1978; Wherry and Bartlett, 1982). The obvious proposition here is that the greater the ability of the rater, the greater the accuracy of rating.

Given the importance of the motivation of the rater (Bernardin & Cardy, 1982), it is surprising that only one study, DeCotiis & Petit (1978), explicitly considered motivation as an independent variable. A rater may be fully prepared to make accurate ratings, but have no incentive to initiate or persist at this task. While the inclusion of this variable certainly strengthened the DeCotiis & Petit model, it was weakened by the failure to include the cognitive processes of the rater. These variables were also excluded by Spool (1978) and Ilgen (1983). As Wherry and Bartlett (1982) noted, an essential part of the rating process is the observation, storage, and retrieval of performance information by the rater. As will be shown in the next chapter, these processes may produce inaccurate ratings because of the limited information processing capabilities of the rater.

Finally, the DeCotiis and Petit, and Bartlett and Wherry models pay careful attention to the contextual factors which may account for variance in performance rating accuracy. In particular, attention is given to organizational policies, composition of the work group, the technology of the work place, training given to raters, and the rating format. The Spool and Ilgen models disregarded these

important features with the exception of the rating format available.

In summary, these models can serve as a guide to future research concerning the accuracy of performance ratings. However, certain additions and refinements must be made to each model. In particular, any model of the rating process must consider the characteristics and performance of the ratee; the values, ability, motivation, and information processing capabilities of the rater; and a number of contextual variables having an impact upon the rater and ratee including organizational policies, composition of the work group, technology of the work place, training given to the raters, and the rating format. Moreover, the role of feedback needs to be incorporated. One might expect, for example, that the inaccuracy of supervisor's ratings would have an impact on the performance of the ratee. The ratee might try to conceal or distort behaviors that are being inaccurately perceived, or bring these inaccurate perceptions to the attention of the rater. In turn, these actions by the ratee may have an impact on the accuracy of future ratings. These and other potential feedback loops need to be incorporated into these models.

Empirical Studies

A small number of empirical studies have been made to test the various components of these models. As will be seen in the review of these studies to be presented here, the majority of them have focused on contextual factors and

ironically, have been conducted in the laboratory. These studies are grouped into the following categories: characteristics of the rater, characteristics of the ratee, and contextual variables. Only those studies that treated performance rating accuracy as the dependent variable were reviewed.

Characteristics of the Rater

Personality. Borman (1979) looked at the relationship between individual difference measures for the rater and their differential accuracy scores. The sample was made up of 146 university students. Individual difference variables were measured using the Minnesota Person Perception Battery. These individual difference measures accounted for 17 percent of the rating accuracy variance. The results suggested the following profile for accurate raters. They tend to be stable, dependable and good-natured persons. Seldom would they be rebellious, arrogant, careless, headstrong, irresponsible, disorderly, or impulsive. In addition they tend to be characterized as even-tempered, outgoing, patient, affiliative, and mature. Finally, they are likely to be informal, pleasant, logical, unselfish, mature, verbally fluent, conversationally facile, and initiators in social relations.

Borman cautioned the reader that these results are based upon low correlations with differential accuracy scores. It should also be pointed out that no theory or rationale was given for including the variables in this

study. Caution is again advised in the interpretation of these results, although this study does seem to point toward a fruitful line of future inquiry.

Memory capacity. An additional individual difference variable was identified by Rush, Phillips, and Lord (1978). They found that the memory capacity of 144 university students was significantly related to the accuracy of the recall of specific events. They found that high memory capacity subjects, as measured by the Picture-Number Test, MA-1, Educational Testing Service, 1962, were more accurate. Hence, high memory capacity should be added to the performance profile of the accurate rater. Given the high recall demands in the performance rating process (Bartlett & Wherry, 1982) this is to be expected.

Values. The effects of the values held by raters on a difference score measure of accuracy were assessed in a laboratory study by Wexley and Youtz (1983). Female Program Aides (W=23) for a service organization participated in this experiment. The Wrightsman (1964) Philosophy of Human Nature Scales provided measures of the following variables: Trustworthiness, independence, altruism, and variability in human nature. After completing these scales the subjects watched the videotaped performance of a supervisor (Heneman & Wexley, 1983) and then rated the supervisor using a frequency of behavior scale.

The results indicated that accuracy was negatively correlated with the raters' beliefs in other peoples'

independence and altruism, and positively correlated with beliefs about their variability. Hence, raters that have strong beliefs in the altruistic and independent nature of man tend to make less accurate ratings, while those who believe in the variability of human nature tend to make more accurate ratings.

General impressions. Nathan and Lord (1983) examined, in a laboratory study with 120 undergraduate subjects, the relationship between the general impression of the lecturer held by the students and the inaccuracy of the student's ratings of the lecturer. The results suggested that general impressions correlated significantly with only some of the different measures of inaccuracy. Hence, it appears that some, but certainly not all, of the incidents recalled are guided by the rater's general impressions. Memory may not always be guided by pre-set categorization schemes, even though this is the prediction made by schematic memory theorists (Alba and Hasher, 1983).

Characteristics of the Ratee

Differential accuracy phenomenon. The effects of the correctness of the behavior observed on accuracy has been studied by Gordon (1970 & 1972). In particular he identified an effect which he labeled the differential accuracy phenomenon (DAP). This concept suggests that correct behavior (i.e. acceptable or desirable behavior) is likely to be identified more accurately than incorrect behavior (i.e. unacceptable or undesirable behavior).

In his first study (Gordon, 1970), he had 118 managers view the videotaped performance of 19 stimulated "agent-prospect" interactions. A number of correct and incorrect behaviors had been built into these tapes. Accuracy was measured as the number of responses where the subject's response matched the correct or incorrect behavior designation in the script. The results indicated that there was a significant main effect for the DAP. Correct behaviors were rated accurately 88 percent of the time while incorrect behavior was accurately rated about 74 percent of the time. Gordon attributed this phenomenon to the idea that raters tend to overlook incorrect behaviors. Perhaps this is because the identification of incorrect behavior may require the rater to engage in an undesirable task; namely, confronting an employee with a performance problem.

His second study (1972) used the same videotapes and true scores, but this time his subjects were 46 senior marketing students. In this study he also used a one item measure of how favorable an impression the ratee created. Subjects were assigned to a favorable and unfavorable condition created by the manipulation of the background data of the ratee given to the raters. The ANOVA results again indicated a significant main effect for the DAP. The accuracy of the correct behaviors was about 89 percent while the accuracy of incorrect behaviors was about 76 percent. This effect accounted for 45 percent of the variance in accuracy! The favorability X DAP effects were nonsignifi-

cant. These results indicate that the favorability of the background data on the ratee does not have an effect on accuracy and that the DAP operates independent of favorability.

Although they did not directly test the DAP, some indirect support was generated by Nathan and Lord (1983). They separated 120 undergraduate psychology students into one condition where the majority of critical incidents exhibited by a lecturer on videotape were examples of correct behavior and another condition where the majority of incidents were examples of incorrect behavior for the same lecturer. A significant main effect was found and raters were more accurate in recalling the number of times and behavior occurred when the majority of behaviors exhibited by the lecturer were correct rather than incorrect.

Leader behavior. Rush et. al. (1981) found that the amount of structured behavior used by the leader of a problem solving group was related to accuracy. Undergraduate subjects (N = 144) watched the videotaped performance of a leader in a problem solving group. Accuracy was defined as the number of times the subjects correctly recalled stimulus information on the tape. In one tape, the leader was coached to exhibit a high degree of structuring behavior and on the other tape was coached to exhibit a low degree. The main effect for this manipulation was significant and subjects gave more accurate ratings for the high structured leader. This result is not surprising given the

sample of college students who were probably more familiar with structured behaviors through their classroom experiences. It does, however, suggest the possibility that the rater's familiarity with the ratee's job is a determinant of accuracy.

Performance feedback. In the Rush et. al. (1981) study previously described, an additional manipulation took place. The subjects were told immediately following the videotape, that the problem solving group they had observed was the second best or second worst of 24 groups performing the task, or were not given any information. This manipulation had a significant main effect on rating accuracy. The direction of this relationship was not presented by the authors, nor was an explanation offered. Hence, further discussion of this finding is not possible.

Contextual Variables

Rater training. A number of studies have looked at the relationship between rater training and performance rating accuracy. Wakeley (1961) conducted two of the original studies in this area. In the first study, 139 undergraduate psychology students were used as subjects. Two measures of rating accuracy were used: accuracy in judging others and ability to judge differences between people. Both measures tested the subject's knowledge of the beliefs and values of the interviewees in a series of four to five minute interviews shown on videotape. Subjects were assigned to six training conditions and a control group; pre and post

measures were taken. Training consisted of a very short lecture. The six training conditions emphasized observing self, observing others, inferring individual differences, looking for similarities with others, rating error reduction, or a combination of these five programs. Relative to the control group, only two training conditions increased accuracy: looking for similarities in others and the combination.

In a second study, 31 evening MBA students were assigned to two training conditions and a control group. One training condition was the previous one emphasizing similarities in others and the second condition was a combination of looking for similarities in others and observing others. Relative to the control group, both training conditions produced significantly higher pre-post accuracy score gains. While this study is illustrative of the variety of programs that might be used to increase accuracy, a number of limitations preclude firm conclusions. The test-retest and internal consistency coefficients were quite low for the criterion measures. The samples were small and consisted of students. Perhaps the major conclusion to be drawn is that lectures are ineffective in increasing rating accuracy.

Borman has conducted two studies in an attempt to increase the accuracy of raters. In the first study, 90 managers in a large, nationwide insurance company served as subjects (Borman, 1975). Pre and post measures of

differential accuracy were taken and no control group was used. Subjects observed the hypothetical performance of a first line supervisor and rated that person using a BES. Training consisted of a five to six minute lecture on halo error. Accuracy was increased for only two of the six BES performance dimensions. Again, the lecture approach to training, with emphasis on halo error this time, was relatively ineffective. When a lecture was used to warn subjects about several rating errors, the same conclusion emerged (Zedeck & Cascio, 1982).

In his second study Borman (1979b) used a different method of training to increase accuracy. College students (N=123) were assigned to a training condition and a no training condition. Subjects in the training condition were given practice and feedback in eliminating rating errors using three hours of the Latham et. al. (1975) training program. The tapes viewed by the subjects consisted of five to nine minute vignettes of the performance of a recruiter and a manager presented in counterbalanced order. Post-test ratings were gathered using four different formats (BES, trait, summated scale, behavior summary), and differential accuracy scores were calculated. The results showed that training had a significant impact on halo error, but did not have a significant impact on accuracy. Hence, at first blush, it appears that not only is a lecture ineffective in increasing accuracy, but so is a sophisticated training program using practice and feedback. An alternative

explanation offered by Fay & Latham (1982) is that the subjects were college students and hence, were not motivated to rate accurately. Another explanation, to be more fully explained in the next chapter, is that the content of the training program was directed toward the elimination of rating errors. Consequently, it is no surprise that this training had an effect on halo error, but not on accuracy.

In another study, Bernardin and Pence (1980) used 72 undergraduate psychology students for subjects. These subjects were assigned to two training conditions and a control group. In the first condition, rater error training, rating errors were defined and illustrated with distributions of scores, and a discussion took place concerning desirable and undesirable distributions of scores. In the second condition, rater accuracy training, the subjects received a lecture concerning the multi-dimensionality of performance and the importance of fair, unbiased, and accurate ratings were emphasized. Discussion then took place concerning the dimensions of performance for a classroom instructor and the subjects generated examples of high, medium, and low behaviors for all dimensions. The subjects observed a videotape of a classroom instructor and then gave ratings with a BES. Difference scores were calculated between subjects scores and scores from untrained, undergraduate students. A post-test only design was used.

The results showed that raters given rater accuracy training or no training were significantly more accurate

than raters given rater error training. No significant differences were found between the rating accuracy training and control group subjects. These results suggest two possible conclusions. First, as with the Borman (1975) study, lectures concerning desirable and undesirable score distributions do not have an effect on accuracy. Second, the accuracy of ratings is likely to be greater when the emphasis is on accuracy rather than errors. However, this second conclusion must be tempered by the fact that a poor method of rater error training was used. Finally, the results of this study are highly suspect given the fact that untrained undergraduate student ratings were used as true scores.

Rater accuracy training was also the focus of a study conducted by Thornton and Zorich (1980). In this study, 170 undergraduate psychology students were assigned to two training conditions and a control group. In the behavioral training condition, the subjects received a lecture where they were told to observe carefully, look for details, take notes, and note specific verbal and nonverbal behaviors. In the error training condition, subjects were lectured on systematic biases in ratings and were given the instructions provided to the behavioral training group. The subjects observed a 45 minute videotape of a leaderless group discussion and made post-test ratings on the occurrence of specific behavioral events. Accuracy was measured as the number of correct responses.

The results of this study indicate that subjects receiving error training were significantly more accurate than subjects receiving behavioral training, and that both groups were significantly more accurate than the control group. Caution must be exercised in interpreting these results as once again, true scores were generated by untrained undergraduate students. Given this limitation, these results suggest that lecture based training can increase the accuracy of ratings when the ratings call for the correct identification of specific behaviors. In addition, the results suggest that training emphasizing the elimination of rating errors and accurate observation is more effective than training only emphasizing the elimination of rating errors.

Pulakos (1983) compared rater error training (RET) and rater accuracy training (RAT) using 108 undergraduate students as subjects. The former program was similar to the one conducted by Latham et. al. (1975) while the latter program used the rating instrument itself as a training tool along with focusing rater attention to the particular job performance dimensions and their corresponding levels of effectiveness. In both cases the rater practiced making ratings and received feedback on the accuracy of their ratings. A completely crossed fixed-factors design was used and consisted of the following conditions: (1) RET alone; (2) RAT alone; (3) RET/RAT together; (4) and no training.

When a distance measure of accuracy was calculated, the results indicated that RAT alone or RAT and RET together yielded ratings with higher accuracy than no training or RET alone. In addition, there was no significant difference between no training and RET alone, and RAT alone led to a significant increase in accuracy for three of the five dimensions on the rating scale. When a correlational measure of accuracy was used, the results indicated that the subjects in the RAT alone condition produced the most accurate ratings. There was no significant difference between the RET alone and the RET/RAT condition; both conditions, however, produced more accurate ratings than the no training condition.

The final study of this type was conducted by Fay and Latham (1982). They assigned 90 business students to a training and no training condition and to three rating format conditions (BES, BOS, and TRAIT). The subjects were given four hours of rater error training using the procedures set forth by Latham et. al. (1975). Accuracy was calculated with respect to halo, contrast, and first - impressions errors using difference scores.

Training led to more accurate ratings than ratings in the control group, regardless of the rating format. These results, at first glance, seem to indicate that rater error training in and of itself increases accuracy, unlike the results from the studies by Borman (1975, & 1977b), Bernardin and Pence (1978), and Pulakos (1983). This would be

expected given that Borman (1979) and Bernardin and Pence (1978) used a lecture method rather than a method incorporating practice and feedback. However, it does not explain why the Borman (1979b) and Pulakos (1983) training programs did not effect accuracy while in this study it did. All three studies used essentially the same training procedures, although it should be noted that the Fay and Latham (1982) program lasted a longer period of time.

In addition to the differences in the amount of training time, two alternative explanations are possible. First, as Fay and Latham indicated, the business students used in their study may be more motivated to rate accurately than the liberal arts students used in the Borman (1979b) and Pulakos (1983) studies. Second, accuracy was defined in different ways. In the Borman (1979b) and Pulakos (1983) studies, accuracy was assessed without any consideration given to rating errors whereas in the Fay and Latham (1982) study, accuracy was defined relative to rating errors. As suggested before, it may be the case here that when rater error training is used to increase the accuracy of overall ratings, it is ineffective; when it is used to diminish rating errors it is effective.

Time delay in rating Several studies have provided evidence that the accuracy of ratings diminishes as a function of the delay between the observation and rating of performance. In one study (Rush, et. al., 1981) 144 college students were placed into an immediate rating condition and

a 48 hour delay rating condition. The subjects rated the videotaped performance of a leader in a problem solving group. Accuracy was measured as the number of times the subjects correctly recalled stimulus information on the tape. The results indicated that subjects giving a rating immediately following observation were more accurate than those subjects giving their rating 48 hours after observation. Similar findings have been reported for a 48 hour delay (Nathan and Lord, 1983) and for a delay of up to three weeks (Heneman and Wexley, 1983). In addition, Rush et. al. (1981) found that this effect was independent of the memory capacity of the subjects and the type of performance feedback about the ratee given to the rater.

These results underscore the importance of cognitive processing in the rating process. In particular they point to the futility of the common practice of having supervisors make ratings on a yearly basis. It should be noted, however, that the findings are from laboratory experiments and they need to be extended to a field setting. In addition, explanation is needed as to why this effect takes place. The nonsignificant results for memory capacity reported by Rush et. al. (1981) suggest that it is not the result of the information processing capacity of the rater. It may then be due to the passage of time itself between observation and the rating or may be due to the distortions that occur within this period of time (Heneman & Wexley,

1983). These and other explanations need to be further explored.

Observation: Amount and method. In a fixed factors design, Heneman and Wexley (1983) manipulated the amount of information observed by 180 undergraduate business students. In the first condition, subjects watched a 55 minute videotape of a production supervisor interacting with his subordinates in a manufacturing exercise. In the second and third conditions, the subjects viewed a random sample of 60 percent and 20 percent of the critical incidents exhibited by the supervisor in the 55 minute tape. This main effect was significant and the subjects ratings were more accurate the greater the amount of information observed. Future research of this type might hold constant time or the number of observations in each condition to see which one is responsible for the amount of information effect.

Maier and Thurber (1968) examined the manner in which information was presented to the rater. They explored three different methods. Undergraduate psychology students (N=219) were asked to decide whether a student did or did not cheat on an exam. They were divided into three groups in which they watched and heard a live role-play of the incident, heard a recording of this same role play on audiotape, or read a transcript of the role play. The authors found that the subjects that read or listened to the role play were significantly more accurate than those that watched and heard the live performance. There was no

difference between the raters that read or listened to the role play. The authors attributed this finding to the fact that the raters in the two most accurate conditions had the opportunity to go back and review what was said. The implication here is that raters may need to make better use of unobtrusive measures of employee performance (e.g. actions described in letters, memos, and reports).

Purpose of the rating. Zedeck and Cascio (1982) found that the purpose of the rating accounted for 19 percent of rating accuracy variance. They assigned 130 undergraduate psychology and business students to three purpose conditions: recommending development, awarding a merit raise, or retaining a probationary employee. The subjects read a 33 paragraph description of the performance of supermarket checkers. Each paragraph contained information on one checker. The results provided the following rank ordering of the ability of raters to discriminate between ratees: retention, development, merit pay. This effect was significant. The implication here is that as the consequences of a decision increase, accuracy decreases. However, it should be noted that accuracy refers to discriminability here and this is a necessary but not sufficient condition for accuracy.

Format and dimensions. In two of the studies just described, the effects of the rating format on rating accuracy were examined. Borman (1979b) found a significant effect for rating format. However, a significant job X format

interaction indicated that there was no one format that was consistently better than the others for both jobs. Fay and Latham (1982) found that raters using behavioral scales (BES and BOS) were more accurate than raters using a trait rating scale. Osburn, Timmreck, and Bigby (1981) demonstrated that specific dimensions relevant to critical job behaviors were used more accurately than generalized job dimensions for 52 experienced interviewers shown simulated job interviews on videotape. Taken together these results indicate that formats with specific and behavioral statements are more likely to be rated accurately.

Borman (1977) also looked at the effects of the dimension being rated on performance rating accuracy. He found that on the whole these effects were consistent across rating formats. It would appear that some dimensions of performance are less ambiguous to the rater than others. Unfortunately Borman did not report which dimensions were most accurately rated.

Rating errors. A disconcerting finding in the literature to date has been the positive relationship observed between halo and accuracy (Borman, 1977 & 1979b); Berman & Kenny, 1976; & Warmke, 1980). This finding runs counter to a classic postulate in psychometric theory that predicts that accuracy decreases as halo increases. One explanation for this finding was tested by Cooper (1981) who suggested that this result may be due to unreliability in the halo and accuracy measures. He took the correlations

reported in these four studies and corrected them for attenuation using conservative estimates of halo and accuracy reliability. Even after these corrections had been made, he found a median correlation of positive .275. Attenuation does not appear to be the answer.

A number of alternative explanations can be offered and need to be researched to resolve this paradox. First, given the small sample size used in these studies, the results may be due to sampling error. Second, there may be restriction in range because of the college samples employed. Third, halo may have been present in the performances viewed by subjects. That is, there may have been valid rather than invalid halo (Bartlett, 1983). A laboratory study conducted by Pulakos (1983) provides some support for this explanation. Finally, training to eliminate halo error when it is in fact not an error, may decrease accuracy (Bernardin & Pence, 1980).

Summary

In summary, only a few of the propositions contained in the models reviewed in the previous section have been tested. The majority of these findings have dealt with contextual variables. The motivation to rate accurately and the cognitive processes involved in rating, two promising avenues of research as will be described in the next chapter, have received little attention thus far. The most robust finding to date has been the differential accuracy phenomenon (Gordon, 1970 & 1972). On the whole, however,

the results have been disappointing. When effect sizes are reported, they seldom exceed .05. Perhaps this is to be expected when contextual variables are studies in a laboratory setting with undergraduate subjects. It appears that more field research is necessary.

True Score Development

As shown in the definitions section of this chapter, it is imperative that there be a "true" score in order to define accuracy. A true score is the correct or actual behavior or performance engaged in by the ratee over time. An accuracy score is developed when the relationship between the true and observed score is calculated. Not only is it quite difficult, if not impossible, to develop a perfect measure of actual employee performance, but even perfect measurement does not guarantee that this true score will be relevant (Thorndike, 1949). That is, this true score may not be related to the ultimate contribution of the employee to the organization. Finally, even if we did develop a perfectly relevant and true score for one employee, it may be one that in terms of research, is not externally valid. That is, the employee and job are not representative of the universe we wish to generalize our findings to. For all these reasons, current research utilizes measures which only approximate true scores and which may or may not be relevant or externally valid. These attempts to develop an approximation to a true score will be briefly discussed in

this section and some suggestions will be offered as to directions to be taken to develop better true scores.[3]

Fay and Latham (1982) developed true scores by using the videotapes constructed by Latham et. al. (1975) and described in detail in the methodology chapter of this dissertation. Briefly, these tapes showed applicants being interviewed for a clerk and management trainee position. True scores were developed by editing the tapes such that sections of the tapes intended to elicit rating errors (e.g. first-impression error) were eliminated. Then, the ratings of these modified tapes provided by 40 upper level business students were used as true scores. Hence, rather than using "experts" to eliminate rating errors in the videotape, the authors physically removed these errors from the tape.

The external validity of these tapes is good as the jobs (clerk and management trainee) and situation (interview) are familiar to most people. In addition, if the purpose is to measure the difference between ratings that are free from ratings errors and ratings that contain these errors, the method of physically removing these errors rather than relying upon experts appears to be a good one. However, it cannot be assumed that the resultant ratings of this tape present perfect true scores. This is still an approximation to a true score as the true scores may be

---

[3]A review of all the studies in which true scores were developed is not presented here. Instead, the studies included in this section are illustrative of the major methods used to develop true scores.

given by experts with inadequate observation skills. In order to deal with this problem, after the judgment errors have been removed from the tapes, the expert raters can be trained in observational skills before making their ratings. In other words, a combination of physically removing errors from the tapes and using expert, trained raters may be the best approach. In the Fay and Latham (1982) study, untrained raters were used.

Videotapes were also used by Borman (1977) to develop true scores. The tapes depicted an interviewing situation and a manager talking with a problem subordinate. Intended true scores were generated using expert judges (their backgrounds were not reported) to estimate true means, standard deviations, and intercorrelations between items. The intraclass correlations of these judgements were .81 and .82 for the recruiter and manager jobs respectively. Scripts were then written to reflect these expert rater scores, and actors were used to tape their acting out of the script. Next, fourteen new expert raters (graduate students in psychology and practicing industrial psychologists) observed the videotapes. Before doing so they reviewed the scripts and took notes while observing the videotapes. These ratings were then used as true scores. The median intraclass correlation for each dimension was .93 and the median correlation between expert ratings and intended true scores for each dimension was .93.

Like the Fay & Latham (1982) videotapes, these tapes depicted well known industrial situations. Through the use of intended true scores a high degree of realism and control over the behaviors exhibited by the actors was made possible. Moreover, the true score ratings had a high level of reliability - a necessary condition for a true score. Again, however, these tapes were not flawless. In particular, each tape was only five to nine minutes long. In most organizations impressions of employee performance are made over much longer periods of time. In addition, the expert raters were not provided with any training for accuracy or rating errors. However, given their positions, it can probably be assumed that they had been exposed to these ideas at some point in their careers.

In another study of accuracy, Bernardin and Pence (1980) used videotapes to construct true scores. The videotapes, developed by Eder, Keaveny, McGramm, and Beatty (1978) depicted critical behaviors exhibited by a classroom instructor. True scores were developed from the ratings of 27 untrained undergraduate students.

There are two flaws associated with this method of true score development. First, the generalizability of the job (classroom instructor) to industrial situations is questionable. Second, and more importantly, the experts that gave ratings which served as true scores did not receive any training or the author's did not report having done so. Hence, the expert ratings of untrained undergraduate students served as the criterion for trained

undergraduate students! A similar problem exists for the true scores developed by Thornton and Zorich (1980).

True scores were also developed by Heneman and Wexley (1983). In this study, 55, 35, and 20 minute videotapes were constructed. The tapes depicted the performance of a production supervisor as he interacted with his subordinates in a manufacturing exercise. Graduate students trained in the critical incident technique (Flanagan, 1954) were used to record the frequency of critical incidents. These counts were used as true scores when at least two of the three experts described the incident in the same way and agreed where it occurred on the tape. While the experts were trained, they were not experienced raters and this may be a limitation to the resultant true scores. Also, only a small number of expert raters were used which creates reliability problems.

In an interesting study by Maier and Thurber (1968) a number of different methods were used. A role-played interview where a student was accused of cheating by his instructor was shown live, tape recorded, and transcribed. In each case, the true score was whether the accused student did or did not admit to cheating. In this case, experts were not needed as an objective true score was possible. While this feature is desirable from a relevance point of view, it is not very similar to a rating situation. Raters are usually asked to make a number of judgements or observe a number of behaviors, not just one. Both the script and the audiotape recording were also an unrealistic depiction

of the rating process as it was not possible for the rater to see the ratee.

The videotapes developed by Gordon (1970 & 1972) and Nathan and Lord (1983) are excellent examples of the manipulation of content in the videotapes. In both studies scripts were developed for the actors which systematically manipulated the favorability (good or bad) of the incidents. Similar efforts could be undertaken to manipulate other factors of performance including the type, frequency, and duration of various behaviors.

In summary, the predominant method of generating true scores is through the viewing of videotaped performance by "experts." Future developers of these scores should use intended true scores, and have an adequate number of trained and experienced raters. These experts should also have the rating format fully explained to them, have scripts of the performances to be observed, and be familiar with or have the job description for the person being rated. If possible or if necessary, rating errors should be edited from the tapes observed by the experts. Finally, to the extent that the results are to be generalized to industrial settings, classroom instructors should not be used as ratees.

While videotapes offer an important element of control for the development of true scores, they suffer from a lack of realism. In particular, they are of short duration, depict simulated work activities, and do not have live performance. In order to counteract these weaknesses, the performance taped should be in actual rather than simulated

situations. It might be possible, for example, to obtain industrial engineering tapes of worker performance. Alternatively, the cameras used to monitor employees and customers in banks and other businesses might be used for this purpose.

Summary and Conclusions

Performance rating accuracy can be defined by either the correlation or the distance between actual employee behaviors and employee behaviors recorded by a rater. A comprehensive model of rating accuracy would include characteristics of the rater and ratee, contextual factors, and feedback loops between these sets of variables. The empirical findings reviewed here suggest that ratings are more likely to be accurate when raters have high memory capacity and are familiar with the ratee's job, and when ratees exhibit desirable or acceptable behaviors. Furthermore, these findings suggest that ratings will be more accurate when the context is such that raters receive training, use a rating format with specific and behavioral items, observe a large number of ratee behaviors, minimize the delay between the observation and recording of performance, and use these ratings to make decisions concerning retention and development rather than merit pay.

As can be seen from this summary the research on performance rating accuracy is very limited. Organizations that wish to increase the accuracy of ratings can take some obvious steps based on these findings, but more attention must be given to this topic if there is to be further

progress in the prediction, understanding, and control of rating accuracy. In particular, a number of revisions need to be made to the models of accuracy reviewed here, and direct tests of the hypotheses deduced from these models need to be made. In order to accomplish this, more field research needs to be conducted and more careful attention needs to be devoted to the construction of true scores. Finally, two promising lines of research, those involving the cognitive processing capability of the rater and the motivation to rate accurately, need to be further developed. These two variables will be the focus in the next chapter.

CHAPTER 3

Models and Hypotheses

In the previous chapter it was pointed out that two of the more important processes in the rating task, the cognitive tasks confronting the rater as he processes performance information and the motivation of the rater to rate accurately, have received very little attention in the literature to date. Drawing upon cognitive processing theories and expectancy theory, the usefulness of these two processes in predicting and understanding rating accuracy will be presented in this chapter. From this presentation, a series of testable propositions will be advanced in this chapter and then, in the next two chapters, a formal test of these propositions will be described.

The major variables and the relationships to be examined are summarized in Figure 1. Circled numbers correspond to the hypotheses in the text.

The information in this chapter is organized in the following fashion. In the first section, a discussion of performance rating accuracy is presented. Within this section a description of cognitive theories of the rating process is presented and is used to generate hypotheses concerning the impact of rater training and rating format on

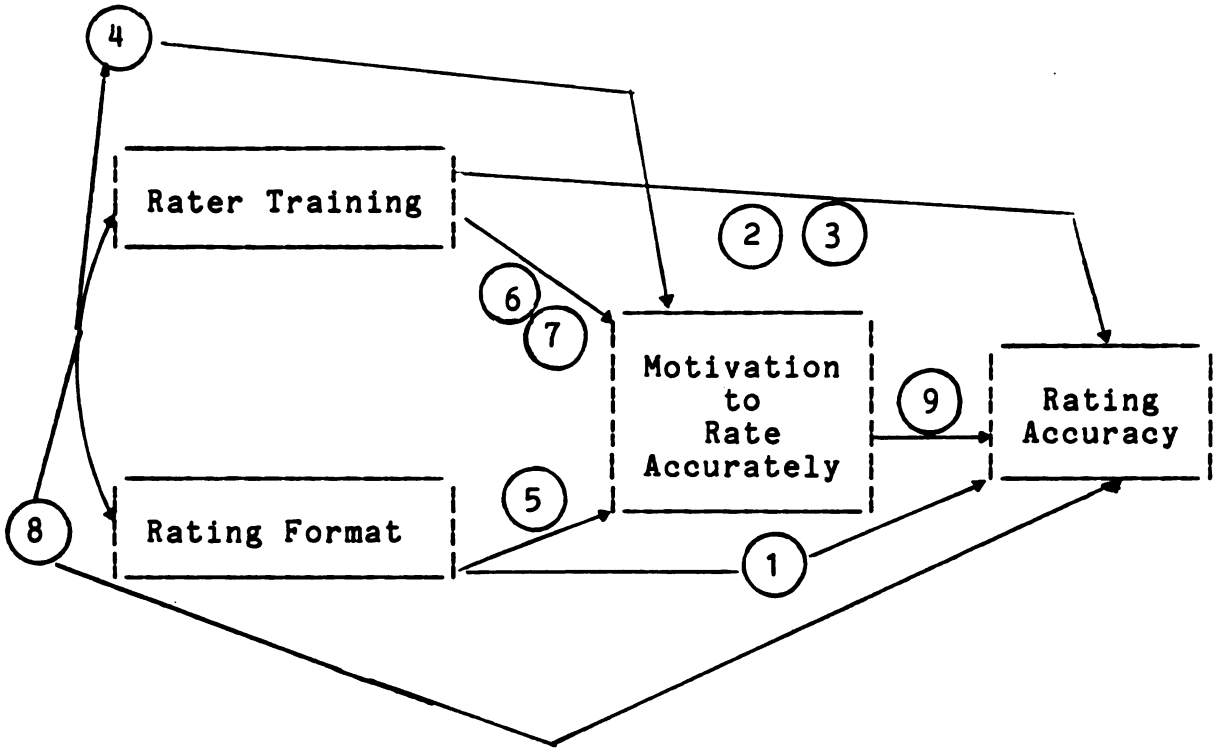

Figure 1. Summary of major variables and relationships.

rating accuracy. The second section looks at the motivation to rate accurately. A description of expectancy theory as it relates to the rating process is presented and is used to make predictions about the effects of rating format and rater training on the motivation to rate accurately. The final section is concerned with the relationship between rating accuracy and the motivation to rate accurately.

Rating Accuracy

Cognitive Processing Theories

A large number of authors have argued that the cognitive processing tasks undertaken by the rater play an important role in the performance rating process and deserve more careful attention (Atkin & Conlon, 1978; Bartlett &

Wherry, 1982; Bernardin & Beatty 1984; Borman, 1978; Carroll & Schneier, 1982; Cooper, 1981; Feldman, 1981; Heneman & Wexley, 1983; Kraiger, 1983; Landy & Farr, 1980 & 1983; Lopez, 1968; Nathan & Lord, 1983; Murphy et. al., 1982; and Wherry, 1952). As a result, several models of this process have been developed. These models will be briefly reviewed here so that they can be used to explain the expected effects of rating format and rater training on rating accuracy.

Wherry (1952), and Lopez (1968) appear to be among the first to describe cognitive processes in the rating process. Wherry (1952), as reported in Wherry and Bartlett (1982), felt that this process involved the observation of performance by the rater and the recall of this performance when a rating was to be made. While not acknowledging Wherry's work, Lopez took this model one step further. He suggested that once the rater had recalled the performance observed, the rater then had to "interpret" or make a summary judgment about the ratee. In addition, he built a feedback loop into this process by suggesting that the recall and interpretation of the performance observed influenced what the rater observed in the next round of observations.

More modern theorists have refined this basic model. In particular, Landy and Farr (1980) pointed out that an additional step takes place between observation and recall. This step is one of storage where the observed performance is organized and integrated with previously stored informa-

tion for recall at a later time. According to Feldman (1981) storage will be an unconscious process when there are existing categories and will be a conscious process when new categories need to be formed to store the observations. Cooper (1981) emphasized that these observations are placed into storage in two phases: short and long term memory. At each stage, distortion in the trace is possible.

Several authors have also refined the judgment stage of the model. Borman (1978) suggested that raters give weights to the various dimensions of performance and then sum up the weighted dimension scores to arrive at a final judgment about the ratee. The importance of attributional processes in the judgment stage have also been incorporated into the model (Bernardin & Beatty, 1984; Landy & Farr, 1983; Carroll & Schneier, 1982). Attributions by the raters concerning the causes and consequences of ratee behavior can influence the rater's judgments of the ratee and the formation of categories for the storage of observations.

In summary, the following components have been included in current models of the cognitive tasks performed by raters as they process performance information: observation, storage, retrieval, and judgment. While these components quite often take place in a sequential fashion, this is not always the case. Feldman (1981) points out that these components are interacting and cyclical. For example, the earlier discussion of Lopez's (1968) work suggested that the retrieval and judgment stages may have an impact on the obser-

vation stage. Given this general overview, two specific issues which have direct implications for the accuracy of ratings will now be addressed.

The first issue concerns the ability of the rater to accurately recall specific, behavioral incidents instead of broad, categorical events. There are two distinct schools of thought with regards to this issue (Alba & Hasher, 1983; Nathan and Lord, 1983). The "traditionalist" viewpoint as represented by the works of Bartlett and Wherry (1982), Borman (1978), and Lopez (1968), suggests that raters are able to store and retrieve the originally observed behaviors. To the extent there are a manageable number of observations to process and the demands on memory are not too great, raters should be able to accurately recall their initial observations and make judgments about the ratee's performance on the basis of these observations.

A decidedly different point of view is taken by categorization or schema theorists as represented by the ideas of Cooper (1981), Feldman (1981) and Murphy, Martin, and Garcia (1982). Because there are so many distinct behavioral observations to process, the ones that are observed, stored, and recalled are based upon predetermined categories or schema developed by the rater through previous experiences. As a result, when ratings are to be made, they are based upon these general categories rather than specific events. In a sense, the specific event is reconstructed from these

general categories and hence, ratings are accurate to the extent this reproduction process is accurate.

At the risk of some oversimplification, these arguments can be extended to predictions concerning the accuracy of two types of rating formats. Borman (1978) has argued that rating formats should be developed to explicitly take into account the cognitive processes of the rater. Given this notion,the traditionalists might contend that a rating format like BOS or BES based upon specific, critical incidents would result in more accurate ratings. On the other hand, schema theorists might argue that global, trait categories more closely approximate the cognitive processes of the rater and are therefore more likely to lead to accurate ratings. These two propositions have not been tested thus far, but serve as stimuli to the hypotheses to be advanced in a latter part of this chapter. Indirect evidence on this issue has been mixed. In a review of the cognitive psychology literature Alba and Hasher (1983) found little emprical evidence and a great deal of theory to support the schema theorists and a large amount of empirical evidence and little theory to support the traditionalist's view. In a laboratory investigation of the performance rating process, Nathan and Lord (1983) found some support for both positions. This issue will be returned to in the discussion of rating format and rater training.

Another issue has to do with whether the observation and judgment stages of cognition are distinct from one

another. One might argue that they are not. As Feldman (1981) points out the four stages of cognition previously discussed are cyclical and thus, observation determines judgment and judgment determines observation. It is also possible to argue that they are relatively distinct. Thornton and Zorich (1975) describe the fundamental differences between observation and judgment in the rating task:

> Prior research in this area has not made clear the distinction between the process of observation and judgment. Judgment processes include the categorization, integration, and evaluation of information. The observation processes are more basic including the detection, perception, and recall or recognition of specific behavioral events, p. 351.

Only one study addresses this issue in the context of the performance rating process. Murphy , Martin, & Garcia (1982) found that the correlation between observation and judgment is modest. In this study, observation was measured using frequency of observation ratings and judgment was measured using a trait rating scale. The four major formulas to measure accuracy, developed by Cronbach (1955), were used. The same raters completed both instruments for the same ratee. Only five of the 16 possible correlations (four accuracy measures using frequency of observations ratings x four accuracy measures using trait ratings) were significant at the .05 level and of these five correlations, the magnitude of the correlation was less than .45 for four of them. Hence, judgment and observation appear to be two separate,

but related concepts, and they will be treated this way in the discussion of rating format and rater training.

Format Effect

Given the cognitive demands placed upon the rater, there are a number of reasons why ratings obtained with the use of a frequency of observation scale should be more accurate than those obtained with a trait rating scale. First, the former format requires less complex judgments to be made (Feldman, 1981). Observation, storage, and retrieval is needed, but the rater does not have to make complex inferences from this information as is true with the latter format (Weick, 1968). As a result, the opportunity for judgmental errors is less frequent (Borman, 1983). Skeptics may argue that raters are poor at recalling specific, behavioral incidents, but the research evidence previously reviewed does not support this contention.

Second, the frequency of behavior scale has an "objective" criterion against which raters can check their ratings (Feldman, 1981). The criterion is simply whether and how often the behavior occurred. It is probably more difficult for the rater to test his ratings when a trait rating scale item like 'dependability' is used.

Third, trait ratings are ambigously worded and interpreted differently be different raters. (Bartlett and Wherry, 1982). Consequently, they are subject to distortion. They are also subject to distortion as they are not directly observable. Frequency of observation items are

directly observable and thus less subject to distortion. Campbell (1961) elaborates upon this point:

> The greater the direct accessibility of the stimuli to the sense receptors, the greater the intersubjective verifiability of the observation. The weaker or more intangible, indirect, or abstract the stimulus attribute, the more observations are subject to distortion, p. 340.

Because traits are not directly observable, inferences must be made from what was observed before a rating can be made (Carroll & Schneier, 1982). These inferences are subject to distortion.

Given these considerations, the first hypothesis to be tested in the dissertation is that:

1. The use of a frequency of behavior scale will produce more accurate ratings than will the use of trait rating scale.

Training Effect

In the review of training programs designed to increase accuracy, presented in the previous chapter, it was pointed out that training which gives raters the opportunity to practice making accurate ratings and to receive feedback on the accuracy of their ratings is more likely to be successful than is training that does not offer practice and feedback. Consequently, the following hypothesis is to be tested:

2. Raters given training that provides practice and feedback on the accuracy of their ratings will be more

accurate than raters that are not given practice and feedback.

It is also expected that the content of the training program will exert an influence on rating accuracy. Present methods of rater training focus on eliminating judgment errors like halo and leniency (Spool, 1978). An alternative type of training focuses on developing the rater's observational skills. To the extent that prototypical behavioral categories (Feldman, 1981) can be established through training and then guide the subsequent processing of performance information by the rater, it is expected that raters will have a more accurate information base from which they can generate more accurate ratings. The observational skills of the rater will be sharpened and also, because there is an obvious carry-over from the observation to judgment stage of cognitive processing (Feldman, 1981), the judgment skills of the rater may also be improved.

Rater error training, which is designed to eliminate judgment errors, is less likely to have as large an impact on both observation and judgment. While there is a carry-over effect from judgment to observation, this effect is probably not as pronounced as the effect of observation on judgment. Before any sort of judgment can be made, some sort of observation must occur. If the initial observation is inaccurate, then the judgment based on that observation is likely to be inaccurate. To the extent this judgment then determines future observations, they are also likely to

be inaccurate even if rater training takes place because of this inaccurate initial observation. As a result, it is hypothesized that:

3. Observational rater training will produce more accurate ratings than will rater error training.

Both of these training programs are described in detail in the methodology chapter of this dissertation. Rater error training appears to be grounded in classic psychometric theory. This theory, as detailed in Bartlett and Wherry (1982), suggests that there are a number of systematic judgment errors (e.g. halo and leniency) that occur as raters process performance information. To the extent these errors can be minimized through training, accuracy should be increased.

The theoretical backdrop for observational training comes from Boice (1983), Flanagan (1944 a, b, & 1952), Flanagan & Burns (1957), and Weick (1968 & 1979). According to Weick (1968) observation is a four stage process. First there is a <u>selection</u> stage where a decision is made about what to observe. This decision is guided by (1) the preset cognitive categories held by the rater (Weick, 1979); (2) the rating scale or some other "standard operating procedure" to which the rater must adhere (Weick, 1979); (3) and the characteristics (Gibson, 1960) and organization (Kohler, 1956) of the stimuli.

Second, there is a <u>provocation</u> stage where the rater must put himself into an appropriate situation to observe

that which has been selected to be observed. The rater is not a passive recepient of environmental stimuli, but instead engages in a process of enactment (Weick, 1979). In doing so, the rater actively creates the environment in which the rater and ratee interact. Thus, the rater is like a "participant-observer" conducting anthropological research (Firth, 1951). Actions taken by the observer have an impact on the ratee's performance.

Third, the rater must mentally or physically record his observations. This is the recording stage. Finally, there is an encoding stage where the rater must mentally or physically keep track of the frequency of similar observations. In order for this process to run smoothly and thus produce accurate ratings, Weick (1968) felt that inferential demands upon the rater must be minimized. Compared to classical psychometric theory then, the objective here is to get raters to minimize the need to make judgments rather than learning how to avoid making errors in judgment.

In order to implement this idea, Boice, Flanagan, and Weick have come up with a number of suggestions which are listed below:

- The selection stage should emphasize categories that are specific enough to be observed, but not so specific that they place unrealistic recall demands upon the rater.
- Categories should guide this rater on what to observe, but should not be some complex that more attention is paid to the categories than

the ratee. That is, performance on the job is equivocal and for it to be captured, the categories must also be equivocal (Weick, 1979).

- Observations should be physically recorded to guard against memory loss. Methods to accomplish this are provided by Flanagan and Burns (1957) and Smith (1982).
- Observations should be directly accessible to the raters senses.
- Raters should put themselves in situations where the behavior to be observed is likely to take place.
- Behaviors rather than traits should be emphasized.
- Incidents that are critical to employee success or failure should be emphasized.

These recommendations are incorporated into the observational training program used in this study.

Format X Training Effect

Given the importance of cognitive processes in the rating task, Borman (1978) has argued that the rating format should be consistent with these processes and Spool (1978) has argued that rater training programs should model these processes. Taking this view one step further, the contention here is that the rating format and rater training program must be consistent with one another.

Frequency of behavior scales and observational rater training focus on the observation stage of cognition, while trait rating scales and rater error training are concerned with the judgment stage. If these two stages are somewhat distinct, and the literature reviewed in an earlier section indicates that they are, then the appropriate matching of format and training should be made. If this is not done, and a training program based upon observation (judgment) is paired with a subsequently used rating format base upon judgment (observation), then the elements of the rater training program are unlikely to transfer to the rating task. Consequently, the following hypothesis is offered:

4. The accuracy of ratings will be greater for those raters receiving observational training and a frequency of behavior scale than for those raters receiving observational training and a trait rating scale. Likewise, the accuracy of ratings will be greater for those raters receiving rater error training and a trait rating scale than for those raters receiving rater error training and a frequency of behavior scale.

Motivation

Expectancy Theory

The importance of the motivation of the rater to rating accuracy has long been recognized (Bayroff, Haggerty, & Rundquist, 1954; Taft, 1955). However, as discussed in the previous chapter, very little attention has been given to

this topic. Two notable exceptions are the models set forth by DeCotiis and Petit (1978) and Mohrman and Lawler (1983). Both of these models take an expectancy theory view (Mitchell, 1974; Vroom, 1964) of the rating process. From this perspective, the decision of the rater to initiate and persist in behavior that will lead to accurate ratings is a function of the belief that effort at the rating task will lead to accurate ratings (expectancy perceptions) and the belief that accurate ratings will lead to certain outcomes (instrumentality perceptions).

The variables discussed by DeCotiis and Petit and Mohrman and Lawler can be set within this context. A major variable presented by DeCotiis and Petit, perceived adequacy of the rating instrument, may have an impact on expectancy perceptions. More specifically, they felt that the motivation to rate accurately was likely to be greater when the instrument is easy to understand and is job related. Similarly, Mohrman and Lawler suggested that the motivation is likely to be greater when an adequate instrument is available and understood. Surprisingly, both sets of authors ignored the role of rater training. One might expect that expectancy perceptions would be higher the greater the skill levels of the rater, assuming that the training is effective.

Several variables advanced by these authors have to do with instrumentality perceptions. In particular, DeCotiis and Petit suggested that motivation will be increased when

the results of the appraisal are confidential from the ratee, when the rater feels that he has the necessary insights into ratee job behavior, when performance is seen as a legitimate role for the rater, and when the purpose of the appraisal is for personnel research or employee development. Mohrman and Lawler also discuss the purpose of the appraisal and go on to look at the perceived consequences of the rating to the rater, ratee, and others in the organization, and finish by examining extrinsic and intrinsic rewards to the rater for accurate appraisals. Some of these variables will now be used to make predictions concerning the impact of rating format and training on the motivation to rate accurately.

Format Effect

There are a number of reasons to expect that the motivation to rate accurately will be lower for raters using a trait rating scale than for those using a frequency of observation scale. First, trait rating scales may be difficult to understand and not job related. Second, as a result of the lack of an objective criterion for the rater to test his rating, the rater may become defensive and not be willing to use the trait rating format (Patten, 1982). Finally, McGregor (1957) suggested that raters are resistent to performance appraisal because they are suspect of the validity of the format and do not like being cast in a judge role. Brumback (1972) suggests that these fears are especially true with a trait rating scale: "... as opposed to

job-oriented scales, person-oriented scales may be more prone to cast supervisors as judges instead of observers, to make them less certain of their ratings, and to be less acceptable to them, p. 569." As a result of these considerations it is hypothesized that:

5. The motivation to rate accurately will be less for raters using a trait rating scale than for those using a frequency of observation scale.

This hypothesis assumes that instrumentality perceptions are held constant. When allowed to vary, this hypothesis may no longer hold. For example, raters may be more motivated when using a trait rating scale because they are held less accountable for their ratings.

Training Effect

Expectancy perceptions should also be strengthened by rater training. Compared to raters receiving no practice or feedback, raters given the opportunity to make ratings and receive feedback on accuracy should have improved skill levels and more self-confidence about these skills (Schneier & Carroll, 1982). Hence, it is hypothesized that:

6. Raters given training that provides practice and feedback on the accuracy of their ratings will be more motivated to rate accurately than raters in a control group that do not receive this practice and feedback.

Similarly, the content of the training program is also expected to have an impact on the motivation to rate accurately. Given that observational training is less complex

(i.e. does not require the rater to learn how to make complex inferences about the ratee's behavior), the expectancy of accurate ratings will probably be stronger for those raters trained in observational techniques. Consequently, it is hypothesized that:

7. Those raters given observational training will be higher in motivation to rate accurately than those given rater error training.

Format X Training Effect

If a training program based upon observation (judgment) is paired with a subsequently used rating format based upon judgment (observation), then two things are likely to happen. First, the elements of the rater training program are unlikely to transfer to the rating task. Second, the expectancy that effort at rating will lead to accurate performance ratings is likely to be diminished. Therefore, the following hypothesis is to be tested:

8. The motivation to rate accurately will be greater for those raters receiving observational training and a frequency of behavior scale than for those raters receiving observational training and a trait rating scale. Likewise, the motivation to rate accurately will be greater for those raters receiving rater error training and a trait rating scale than for those raters receiving rater error training and a frequency of behavior scale.

Motivation and Rating Accuracy

In the previous chapter it was shown that rating accuracy is not only a function of the rater's ability, but is also a function of the rater's motivation to rate accurately. Consequently, it is expected that:

9. There will be a positive correlation between rating accuracy and the motivation to rate accurately.

Causation should not be inferred from this hypothesis. As suggested above, it may be the case that motivation causes accuracy. On the other hand, it is equally likely that accuracy causes motivation (Johnson, 1945). That is, if the ratings are perceived by the rater to be accurate, this may lead to more confidence in the ratings, and in turn this confidence may increase motivation.

Summary

A summary of the hypotheses presented in this chapter is listed below:

1. The use of a frequency of behavior scale will produce more accurate ratings than will the use of a trait rating scale.
2. Raters given training that provides practice and feedback on the accuracy of their ratings will be more accurate than raters not receiving this practice and feedback.
3. Observational rater training will produce more accurate ratings than will rater error training.
4. The accuracy of ratings will be greater when the rater training and the rating format are consistent with one another.
5. Motivation to rate accurately will be less for raters using a trait rating scale than for those using a frequency of behavior scale.

6. Raters given training that provides practice and feedback on the accuracy of their ratings will be more motivated to rate accurately than raters not receiving this practice and feedback.

7. Raters given observational training will be more motivated to rate accurately than those given rater error training.

8. The motivation to rate accurately will be greater when the rater training and rating format are consistent with one another.

9. There will be a positive correlation between rating accuracy and the motivation to rate accurately.

The research methodology used to test these hypotheses is presented in the next chapter. The chapter following the next one presents the empirical results.

CHAPTER 4

Research Methodology

The research methodology used to test the hypotheses generated in the previous chapter are presented here. Two independent variables were manipulated in a laboratory experiment: the content of the training program and the type of rating format. After training, subjects in the four experimental conditions and the two control groups observed and then rated the videotaped performance of a production supervisor managing two subordinates during a manufacturing exercise. These scores were then used to measure the first dependent variable, rating accuracy. Subjects also completed an instrument designed to measure the second dependent variable, motivation to rate accurately. The following sections in this chapter describe the experimental design, manipulations to the independent variables, measurements made, subjects and procedures used, and the method of data analysis.

Experimental Design

A 2 x 3 factorial design was used and is presented in Figure 2.

Rater Training	Rating Format: Traits	Rating Format: Behaviors
Error	Experimental Group 1	Experimental Group 2
Observation	Experimental Group 3	Experimental Group 4
Control	Control Group 1	Control Group 2

Figure 2. Experimental design.

The first factor, rating format, consisted of two levels: frequency of behavior scale and trait rating scale. The second factor, rater training, was defined by three levels: rater error training, observational training, and control group training. Dependent variables were measured after the treatments had been applied and not before. Hence, there were no repeated measures and only a post-test was made. A pre-test was not made in order to prevent any sensitization of the subjects to the test prior to the treatments.

Manipulations

In this section a description of each of the independent variables, training content and rating format, will be presented.

Training Content

Three different types of training were given - rater error training, observational training, and control group

training. A comparison between the different types of training and a description of each one will now be covered.

Comparison of the three training programs. The content or "what was presented" was different for each training program. In rater error training emphasis was placed upon common judgment errors that can occur in the rating process (e.g. halo, leniency) and ways to eliminate them. Observational training focused on a method to establish prototypical rating categories (e.g. critical behavioral incidents). While the emphasis was different for these two training programs, there was one area of overlap. Both programs encouraged the participants to focus on job related rather than non-job related behaviors. Training given to the control groups did not cover rating errors, nor did it cover observation methods. Instead, a general overview of the rating process was presented and included the definition of performance rating, uses for performance ratings, major court cases, and the performance appraisal interview.

The process or "how the information was presented" had some similarities and differences between rating programs. All three of the training programs lasted approximately three hours. Unlike the two experimental conditions (rater error training and observational training), lecture, discussion, and role playing were used to present the material to the control groups. The training method used for both rater error training and observational training was set forth by Wexley, Sanders, and Yukl (1975). It has been labeled the most "advanced" rater training program developed

(Borman, 1979b) as it is heavily based upon learning theory. Emphasis is placed upon giving raters the opportunity to observe and practice performance rating using videotapes, providing raters feedback on the accuracy of their ratings, and making the program meaningful by using realistic stimuli (Spool, 1978). These procedures and the same videotapes were used for rater error training and observational training.

All three training programs were pre-tested by the trainer using groups of 5-10 undergraduate and graduate students. The training was conducted, reactions and points of confusion were elicited, and steps were taken to correct any deficiencies. The training programs presented in this section are the final versions after corrections warranted by the pre-test were made.

In summary, the three training programs were all the same length. The control group differed from the experimental groups in terms of the content of the training program and the process used to present the content. An identical process was used to present rater error training and observational training. The content of these two training programs, however, were quite different.

Rater error training. Subjects in this experimental condition participated in three exercises. In each exercise, subjects watched a videotape of a hypothetical job candidate being evaluated by an interviewer (Latham & Wexley, 1981). Job applicants were applying for a bookkeeper position and a management trainee position. These jobs and

the interviewing situation were chosen for this study because most of the subjects were familiar with these jobs and were responsible for conducting interviews.

Subjects were asked, in each exercise, to rate the performance of the job applicant on nine point scales. Three points on the scale were anchored with a verbal description: 9 - "would recommend strongly that an offer be made; applicant shows excellent qualifications in all areas;" 5 - "would recommend with reservations that an offer be made; applicant has weak qualifications is some areas;" and 1 - "would recommend that no offer be made; applicant obviously unqualified."

The first exercise was concerned with "first impressions error". This error occurs when the evaluation is primarily based upon the rater's initial reactions to the ratee (Latham & Wexley, 1981). Subjects were first provided with and asked to read a job description and a list of the minimum qualifications necessary for the job of bookkeeper. Next, they watched a videotape of a female applicant as she was interviewed for the bookkeeper job. In the beginning of the videotape, the applicant exhibited some unfavorable characteristics that were not related to the minimum qualifications for the job (e.g. dropped her check book, uncertain of the name of the position she was applying for, etc.). In the last part of the videotape, the applicant made it clear that she did have the necessary qualifications for the job (e.g. had appropriate degree and experience).

Subjects were then asked to indicate whether they would hire her for the job using the 9 point scale previously described. After making the rating, the subjects shared their ratings and the reasons for their rating with the rest of the group. At the end of this stage, the trainer gave them feedback on the accuracy of their individual ratings. It was pointed out that a low rating indicated first impressions error and this term was defined for the subjects.

In the final part of this first exercise, the trainer elicited examples of first impression error back on the job from the group. Typical examples included the following:

- The performance of a new employee was rated on the basis of his first few days on the job.
- An employee was constantly given difficult and dirty assignments because his performance on a new task was low. As a result, his performance was rated low on all tasks.

The trainer and the group then discussed ways to eliminate this error back on the job. Solutions discussed included the following:

- Snap judgments should not be made concerning the performance of an employee. Judgments should not be made until the end of the appraisal period.
- Do not make work assignments on the basis of initial impressions of the employee.

The second exercise focused on the "similar-to-me" effect (Latham and Wexley, 1981). This error occurs when a ratee is evaluated more favorably because he is similar to

the rater along non-job related dimensions. Subjects were again asked to read a job description and the minimum qualifications for the job. This time, however, they were for the job of management trainee. The subjects then watched a videotape of a male applicant being interviewed for this position. The videotape showed that there was a high degree of attitudinal and biographical similarity between the interviewer and the interviewee. The interviewee did not, however, meet the minimum qualifications for the job.

After watching this videotape, each subject was asked to make a rating of the applicant using the same nine point scale. They were also asked to indicate, using this scale, the rating they felt the interviewer would give to the applicant. Each subject then presented his ratings to the rest of the group and gave reasons for making them.

At this point, the last part of the videotape was shown where the interviewer committed similar-to-me error by telling the next person to interview the applicant, that this applicant was an excellent candidate because of a similar attitudinal and biographical background. The trainer then gave the subjects feedback on the accuracy of their ratings. It was suggested by the trainer that subjects with high scores for the applicant had committed similar-to-me error, as verified by the justification they gave for their rating, and this term was defined.

As a final part of this exercise, the trainer elicited examples of similar-to-me error back on the job. The following items are examples of this discussion:

- The children and spouses of the rater and ratee know and interact with each other.
- The rater and ratee ride together in the same car pool.
- The rater and ratee like to get to work early in order to have a cup of coffee together.
- They both like the same sports teams.

The trainer and the subjects then generated a list of ways to overcome this error back on the job. Examples of these solutions are as follows:

- Establish performance criteria for the job before making a rating.
- The rater should check on his ratings by having other raters with a background and attitudes different than his own review his ratings.
- Employees should be rated on how well they perform the job instead of how similar they are to the rater.

The third and final exercise centered around "halo" and "leniency" errors (Latham & Wexley, 1981). Halo error occurs when the rating is based upon someone else's opinion or when one dimension of the employee's performance is generalized to all other dimensions. Leniency occurs when the employee's performance is judged as being high along all dimensions or low along all dimensions when in fact the employee's performance is high on some dimensions and low on others.

The job description, minimum qualifications, and applicants for this job were the same as those used for the previous videotape. This time, however, the interviewer was different. He was the person that came in at the end of the second videotape and heard the previous interviewer rave about how good the applicant was. The new interviewer was impressed with the applicant even though the applicant did not have the necessary qualifications for the job. After watching the videotape, the subjects used the same nine point scale to indicate whether they would hire the applicant and whether they thought that the interviewer would hire the applicant.

Once the ratings had been made, and the subjects explained their reasons for their ratings, the trainer explained to the subjects that a high rating indicated they had fallen victim to halo error and leniency error like the interviewer in the videotape. Both of these terms were then defined by the trainer. Following this, the trainer asked for and discussed examples of halo error and leniency error back on the job. These are some examples of halo and leniency error that were discussed:

- The department is doing poorly because of a lack of of supervision. The supervisor gives all the employees low ratings so he won't look bad to his boss.
- The engineer is good in the technical matters of the job and is therefore rated high for his managerial responsibilities.

- The rater fails to spend time with the ratee and makes his ratings on the basis of what he hears from others.

Finally, the trainer and the subjects generated a list of ways to eliminate this error. Some of these methods are presented here:

- Performance is multidimensional and it is possible for an employee to be high on one dimension and low on another.
- The rater should make his own rating before listening to others that have evaluated the employee.
- The rater should keep notes on what the scale values of the rating scale mean to him.

Observational training. The materials used to conduct this training program were identical to those used in the rater error training program. The same videotapes were shown and in the same order. In addition, the job descriptions, minimum qualifications, and the nine point rating scale remained the same.

The steps taken to present the materials were also the same for each exercise. First, the subjects read the job description and minimum qualifications. Second, each subject rated the applicant, presented their ratings to the trainer and the rest of the group, and gave their reasons for making these ratings. Third, the trainer then gave the participants feedback on the accuracy of their ratings and discussed with the subjects why high or low ratings were given. Fourth, the trainer elicited similar examples that

the subjects encountered back on their jobs and discussed ways to make more accurate ratings.

In the first exercise, subjects were encouraged to look for behaviors rather than traits when observing performance on the job. A behavior was defined to the subjects as an observable activity that the employee engages in while working at the job. Traits were defined as those personal characteristics of the employee that may, but usually are not, related to job performance. The trainer told the subjects, and their discussion of the ratings they gave verified this, that low ratings of the applicant usually indicated that the subject was looking at traits of the applicant (e.g. she was clumsy, awkward, uncertain, etc.). A high rating for the applicant, which she merited, indicated that the subject was focusing on behaviors (e.g. she prepared monthly statements and income tax for her husband and brother).

The subjects were also asked to provide examples of where traits rather than behavior were rated back on the job. Typical traits described were: attitude, appearance, and intelligence. A number of critical behaviors back on the job were also brought forth. For instance, the trainer asked the subjects to define a "bad attitude." Responses varied from refusal to carry out an order to high absenteeism. The trainer emphasized that the focus of performance observation should be on these specific and observable behaviors rather than a vague and unreliable category like "attitude."

Finally, the subjects and the trainer generated a list of ways to make sure that behaviors rather than traits were considered when rating employees back on the job. Typical solutions included the following:

- Review performance records before making a rating (e.g. attendance, safety, and output records.)
- Develop behavioral performance standards for the job.
- Keep in mind Title VII of the 1964 Civil Rights Act.

In the second exercise, the emphasis was on making ratings based upon behaviors rather than noise factors. Noise was defined as those things that the employee says or does that have little to do with performance, good or bad, on the job. When the subjects were asked to give the reasons for their high ratings they usually said that it was because the applicant was a "nice person", and when pushed a little further it was because the applicant and interviewer seemed to have so many thing in common (e.g. lived in same part of town). It was emphasized that these were noise factors rather than behaviors that they were paying attention to.

Subjects were again asked to think back to their own jobs and come up with examples of noise factors that should not be attended to for performance rating purposes. Examples included giving someone a low rating because they had poor table manners and giving an employee a high rating because they came in early to prepare coffee for the boss.

The subjects were asked to generate a list of ways to insure that attention was given to behaviors rather than noise factors. Typical solutions included the following:

- The rater should not overemphasize or encourage discussing of personal matters on the job. Personal problems could be referred to the Employee Assistance Program.
- Performance requirements for the job should be job related and behavioral.

In the third and final exercise, raters were encouraged to consider critical rather than non-critical behaviors when observing performance. A critical behavior was defined as one that produced should good or bad results that the rater wished that every employee would do it all of the time or never do it. A noncritical behavior was defined as one that is routinely expected of and done by the employee. In the videotape, both the applicant and the interviewer exhibited a number of critical behaviors. For instance, the applicant had not completed his college degree which was required in the minimum qualifications. On the other hand, the interviewer asked very leading questions where the answer he sought was given in the question. To the extent these critical behaviors and others were overlooked by the subjects, as witnessed by the reasons they gave for their ratings, they tended to mistakenly give high ratings to the applicant.

After this point was made by the trainer, the subjects were asked to think of examples of critical and noncritical

behaviors back on the job. Typical examples are as follows:

- It is one thing for an employee to fill out a report, but it is critical when it is done accurately and in a timely manner.
- When setting up improvement plans with an employee it is critical that supervisor follow up on these plans and provide assistance if it is needed.

The trainer also asked and helped generate ways to focus the raters attention on critical behaviors back on the job. Some of the solutions are presented below:

- Actively seek situations where the employee is likely to be or should be engaged in critical behaviors. This does not mean, however, that the rater should be constantly there in those situations as this may carry a message to the employee that he is not trusted and make the employee resentful (Purcell, 1955).
- Focus on behaviors and outcomes rather than activities.
- Use the critical incident technique (Flanagan, 1954). Look for the situation, the observable activity that the employee was engaged in, and whether it had large consequences for the company, peers, the customer, etc.

Control group training. The training provided to the control groups did not present any material on rating errors or observation, nor did it make use of the videotaped performance of employees. Instead, a general overview of the

performance rating process was presented using a combination of lecture, discussion, and role playing.

In the first phase of the workshop, the trainer lectured on the definition of performance ratings and presented them with the results of a survey conducted by Downs and Mocinsky (1979). The survey listed the frequency of use for various performance rating systems in Fortune 500 companies. Also in this first phase, the trainer lectured on what could be rated, traits, behaviors, and results, and provided them with examples of each one.

In the second phase, the trainer spoke about the use of performance ratings for the practicing supervisor. The discussion centered around merit pay, promotion, feedback, interviewing, diagnoses of performance problems, training and coaching. Then the trainer lectured on the characteristics of a good performance appraisal. Emphasis was placed upon reliability, validity, fairness, discriminability, and practicality (Latham and Wexley, 1981). Finally, significant court cases concerning performance appraisal were presented by the instructor. In particular, factors which determined whether the court sided with the plaintiff or the defendant were reviewed, based upon an empirical examination of 66 court cases by Feild and Holley (1982).

At the end of this second phase of the training, a general discussion took place concerning two performance rating forms: a graphic rating scale and a MBO type plan used by their organization. They were asked to compare and contrast each form in terms of how well they met the cri-

teria for a good performance rating system, how well they met the criteria for a legally defensible performance system, and ways that the supervisor could use them.

In the third and final phase of the control group training, the trainer lectured on performance appraisal feedback and goal setting. In particular, emphasis was placed on various ways to conduct a feedback session and when each approach might be appropriate (Wexley, 1982). In addition, the trainer reviewed a list of critical incidents concerning effective and ineffective behaviors when giving performance feedback (Latham & Wexley, 1981) and reviewed specific techniques for setting goals (Locke, Shaw, Saari, & Latham, 1981). Finally, subjects were placed in groups of three, given a completed appraisal form, and asked to practice giving feedback and setting goals by role playing the rater and ratee depicted on the appraisal form. The trainer went from group to group and gave the subjects feedback. At the end of the role playing exercise, a general discussion took place.

It should be emphasized that both the content presented and the process used for the control group training were different than the training given to the experimental conditions. Videotapes were not used and rating errors and observational techniques were not discussed. The decision to provide the control groups with a treatment was made to eliminate a rival explanation to the data. To the extent that the motivation to rate accurately was higher for subjects in the experimental conditions than in the control

groups, and the control groups received no training, one could argue that the results did not support the hypotheses. Motivation to rate accurately would not be the result of differences in the content of the training program or the type of rating format, but instead would reflect the presence or absence of training. By providing the control group with training, this alternative explanation to the data was minimized.

Rating Format

Two different types of rating scales were developed for the videotaped performance of a production supervisor. A description of the videotape and the two rating scales is presented here.

Videotape description. A manufacturing exercise (Wexley & Jaffee, 1970; Wexley and Nemeroff, 1975) was videotaped (Heneman & Wexley, 1983). This 55 minute exercise required a supervisor and two subordinates to organize and run their business so as to maximize their profit. This work team purchased parts from a supplier, assembled the finished products (i.e. shipping containers) according to specifications, and sold them to a purchaser. During the exercise, the cost of parts and the prices of the finished products varied from one time period to another, thereby changing the margin of profit.

A male manager from a small manufacturing company played the role of the supervisor while two male graduate students served as subordinates. Two weeks in advance of

the videotaping these actors were provided with instructions for the exercise. In addition, the graduate students were given a list of two behaviors to be exhibited at any time they felt it was appropriate during the exercise. The supervisor was given a list of 15 behaviors that could be used during the exercise and was told to use these or any other behaviors that fit his own style of supervision. He was told to evenly distribute his behavior over time during the exercise. No other special instructions were provided to any of the actors, except that they were encouraged to act as they normally would.

Immediately before the taping of the session the instructions to the exercise were reviewed and the actors practiced their assigned behaviors.

Frequency of behavior scale. Three graduate students in organizational behavior reviewed the videotape previously described and two additional videotapes of the same exercise where the supervisor worked with different subordinates. These raters recorded the critical incidents exhibited by the supervisor (Flanagan, 1954). Before making these judgments, the three raters were trained in this process. The definition of a critical incident was reviewed, examples were provided, and practice and feedback in making these judgments were given. The instructions to the manufacturing exercise were also reviewed, as were the job duties of the supervisor to be observed. The end product was a list of critical incidents for each rater based on the performance of the production supervisor in each of the three

videotapes. The order of the critical incidents was also known as the raters kept track of the time that each one occurred.

A frequency of behavior scale was developed directly from these critical incidents and is shown in Appendix A. Items on this scale are those critical incidents where at least two of the three raters described the critical incident the same way and where at least two of the three raters agreed that it occurred at the same time(s) on the videotapes. Each item was anchored with a scale from 0 to 4, which represented the number of times that the item was observed. In summary, the frequency of observation rating scale consisted of 21 items describing critical incidents that could be observed on the videotapes. Fifteen of these incidents occurred at least one time on the one videotape used in this study. Each item had a rating scale ranging from 0 to 4 representing the number of times that the respondent observed the occurrence of that critical incident on the part of the supervisor.

Trait rating scale. Uhrbrock (1961) developed 2000 scaled items to be used for performance ratings. These items were used to develop the trait rating scale. The author went through this list and eliminated those items that did not pertain to the videotape performance of the production supervisor or were not described as traits. More specifically, items were eliminated for the following reasons:

- The same item appeared elsewhere on the list.
- The item was written in behavioral terms (e.g. "Generates, ideas concerning new work methods").
- A knowledge of results was required (e.g. "Consistently exceeds production standards").
- The employee was compared to other employees (e.g. "Is superior to general run of employees").
- The item related to promotion rather than present performance (e.g. "Ready to be promoted at the earliest opportunity").
- Information about the employee was not available on the videotape (e.g. "Meets new people easily").
- Knowledge of the employee's life outside of work was required (e.g. "Has normal home life").
- Information about the background of the employee was required (e.g. "Has good experience for present job").

Items were also eliminated from the list when the scale values indicated low interrater agreement. Uhrbrock (1961) used 160 student and professional raters to develop scale values for each item. These raters sorted the items into eleven piles, ranging from "favorable" to "unfavorable", to form a Thurstone scale. The mean and standard deviation was reported for each item. Those items with a low standard deviation, less than 1.0, were treated here as having low interrater agreement and therefore eliminated.

Using these procedures, the number of items was reduced from 2000 to 187. This new list of items was then presented

to three graduate students in organizational behavior. They were asked to indicate the degree to which each of the 87 items characterized the performance of the production supervisor on the videotape. Each item had a 5 point Likert-type scale benchmarked from "strongly agree" to "strongly disagree."

Before rating the performance of the production supervisor using this scale, the instructions to the manufacturing exercise were reviewed as were the job duties of the supervisor to be observed. In addition, the graduate students were given the rater error training described n a latter part of this chapter.

Items were retained when all three of the graduate students gave an identical rating for that item. Twenty items met this criterion and were used for the trait rating scale shown in Appendix B. In summary, the trait rating scale was made up of 20 items that depicted traits associated with the performance of the production supervisor. Each item had a 5 point Likert-type scale with written descriptions ranging from "strongly agree" to "strongly disagree" attached to each point.

Measures

Four sets of measures were developed for this study: rating accuracy, motivation to rate accurately, reactions, and demographic characteristics. Each one is described in turn.

Rating Accuracy

A rating accuracy score was calculated for both the frequency of behavior scale and the trait rating scale using the following procedures. The Director of Staffing, who was responsible for the performance appraisal system at the organization providing the sample, nominated seven experts in the use of performance appraisal at the company. All seven experts were in the personnel department and, like the subjects, they all had subordinates reporting to them. They were unlike the subjects in that they had all received extensive training in performance appraisal prior to the experiment. These experts were blind to the experiment, but were informed that they were needed to make ratings of the videotaped performance of a production supervisor and that these ratings would be used to evaluate a workshop on performance appraisal being conducted at their organization. Each expert received either observational training or rater error training. After receiving the training, three of these experts rated the performance of the production supervisor using the trait rating scale and four experts rated the supervisor using the frequency of behavior scale. As with subjects in the experiment, the wording of items and scale values were reviewed with them prior to observing the videotape. Unlike subjects in the experiment, however, they were asked to take careful notes as they observed the videotape.

In order to assess the adequacy of the experts scores an analysis of interrater agreement was conducted.

Agreement was defined as the number of items where the distance between each expert's score was two scale points or less away from the other experts' scores for that item. Items that did not meet this criterion were not used in subsequent analyses (items 1, 2, 6, 12, and 16 for the freqeuncy of observation scale and items 4, 13, 18, and 19 for the trait rating scale). As a result, there was perfect interrater agreement for each item given the criterion. Moreover, a nonparametric ω^2 test (Lawlis & Lu, 1972), recommended by Tinsley and Weiss (1975), revealed that the interrater agreement is greater than the agreement expected on the basis of chance for both the frequency of observation scale (ω^2=30.50, $p<.001$) and the trait rating scale (ω^2=14.90, $p<.001$).

The mean expert scores were then used as true scores to calculate rating accuracy. A formula similar to the ones used by Bernardin and Pence (1980) and Heneman and Wexley (1983) was used. The scoring formula used in the present study is presented in equation 2.

$$\left(\sum_{n=1}^{N} D \right) \Big/ N \qquad (2)$$

where:

N = number of items

D = absolute distance of observed score from true score

Prior to making these calculations, several of the items were reverse scored.

Motivation to Rate Accurately

The author was unable to locate a motivation to rate accurately scale in the published literature. Consequently a new eight item measure was constructed and is presented in Appendix D. Items were worded to reflect the degree to which effort at the rating task was perceived by the subjects as leading to accurate ratings and, in turn, if accurate ratings lead to outcomes of value to the subjects. Expectancy perceptions were measured using items 1, 3, 4, 7, and 8, and instrumentality perceptions were measured using items 2, 5, 6. Each item was anchored with a 5 point Likert-type scale and the points were benchmarked with written descriptions ranging from "strongly agree" to "strongly disagree".

Reactions

A three item measure was constructed to assess the subjects reactions to the materials presented. This measure was developed so that it could be ascertained whether differences between conditions were due to affective reactions to the workshop rather than the treatment effects. These items are shown in Appendix E. Subjects were asked to indicate their reactions to what was presented and how it was presented using a 5 point Likert-type scale with written descriptions ranging from "poor" to "excellent" attached to each point. They were also asked to indicate whether they would recommend that other supervisors in their company attend the workshop and responded using a 5 point Likert-

type scale ranging from "strongly disagree" to "strongly agree."

Demographic Characteristics

A final form was constructed to see to what extent the results were due to the demographic characteristics of the sample rather than the treatment effects. This form is shown in Appendix F. Subjects were asked to indicate their age, sex, educational level, position, the number of subordinates reporting to them, the number of subordinates they rated, whether they had received company training in performance appraisal, and their department and geographic location.

Subjects

Sampling Procedures

The subjects in this experiment were 87 supervisors and managers from a western utility company. They were sampled from the population of supervisors and managers for the organization using the following procedure. A cross section of the various departments (e.g. gas and electric) and geographic divisions was taken. Department heads were asked if they would be willing to let their supervisors participate in this project. If the answer was affirmative, then the supervisors in that department were asked if they would be willing to participate in this project. If the answer was yes, then they were included in the sample. In all stages of this procedure, individuals were told that the project consisted of some performance appraisal training.

All those involved were "blind" to the purpose of the experiment and the experimental design being used. Subjects were eliminated from the sample if they did not have formal responsibility for the supervision of at least one employee or if they had never completed a performance appraisal form for at least one of their employees.

Sample Characteristics

Demographic characteristics of the sample are summarized in this section. The sample consisted of 66 males and 19 females (there were two missing values) with a mean age of 40.73. There were 34 different job titles with the titles Administrative Supervisor and Supervising Engineer being the most frequent. The median number of employees supervised by the subjects was 11.78 and the median number of employees rated by the subjects was 5.86. Approximately 52 percent of the subjects were from the companies headquarters and approximately 48 percent of the subjects worked in one of the companies 4 largest divisions. The subjects came from 32 different departments with the most frequent representation from personnel, engineering, customer services, and gas. The modal education category was "some college, no degree" and approximately 77 percent of the subjects had received some type of performance appraisal training from the company prior to the experiment. Finally, the mean number of years tenure with the company was 6.19.

Procedures

After the sample was selected, subjects were assigned to one of the experimental conditions or control groups. The subjects were given a list of alternative dates and locations for the training sessions. They were asked to indicate which sessions they would be available for, and were then randomly assigned to one of the sessions that they could attend. After the six groups had been formed, the experimental conditions and control groups were randomly assigned to these groups.

Outside of differences in the content of the training presented, subjects receiving rater error training or observational training were treated in the same way. The trainer (author of the dissertation) first introduced himself and presented the major objective of the workshop: "To provide _____ supervisors with some modern, proven, and practical techniques to make more accurate performance ratings." In addition, the trainer defined what was meant by a performance rating emphasizing that it referred to merit pay ratings, performance review, and promotion review at the subject's organization.

Second, the trainer defined the term "accuracy" in the context of performance ratings, explained why accuracy is important, and explained what could be done to make more accurate ratings. It was pointed out by the trainer that accuracy was crucial for the acceptance of performance feedback and for "fair" personnel decisions. The point was also made that performance ratings could and had been made

more accurate in other places using training with the videotapes to be presented. The subjects were then given a brief overview of the workshop and the trainer told them that it was important that they contribute and share their ideas and experiences with the trainer and the rest of the subjects.

Third, the trainer presented the reasons for conducting the workshop. The subjects were told that they could expect to develop some new skills in the performance rating area, that their organization was interested in learning if this training program would increase the accuracy of their supervisor's performance rating accuracy, and that the trainer would be using the results of this training program for his dissertation. Finally, the trainer solicited and answered any questions, issues, or concerns held by the subjects. After they had been answered, the subjects introduced themselves to the trainer and the rest of the subjects.

Essentially the same procedure was followed for the two groups receiving control group training. There were, however, three important differences. First, the objective of the workshop was "To provide ____ supervisors with a working knowledge of the performance rating process." No mention was made of rating accuracy as the objective. Second, the subjects were informed that the workshop would be lecture, discussion, and role playing. The use of videotapes in performance rating training was not discussed. Finally, the trainer emphasized that the subjects would gain a better understanding of the rating process. Nothing was said about

increasing their rating skills so that accuracy would increase.

After the training had been conducted, the subjects in all experimental conditions and control groups were told that the final stage of the workshop would be to have them observe the performance of a supervisor on videotape and to fill out some questionnaires concerning the performance of that supervisor and their feelings about the workshop. Next, the subjects were given a handout which explained the manufacturing exercise and job duties of the production supervisor they were about to observe on videotape. After reading these descriptions, the trainer answered questions concerning the exercise and the job duties of the supervisor.

The trainer then put the rating scale they would be using on the overhead projector. Subjects in the frequency of observation scale condition saw the frequency of observation scale and subjects in the trait rating scale condition saw the trait rating scale. Both of these scales were described in an earlier section of this chapter. In each case the trainer told them that after viewing the videotape they would use this scale to rate the performance of production supervisor, explained the scale, asked them to carefully read each item, and answered any questions concerning the wording of items or benchmark descriptions on the scale. Immediately before showing the videotape, the subjects were asked to pay careful attention to the production supervisor and not to take any notes during or after the videotape.

This latter instruction was issued so that the results did not reflect note taking behaviors by the subjects rather than the training content or rating scale treatments.

Immediately following the viewing of the videotape the trainer passed out a consent form that had been approved by the University Committee on Research Involving Human Subjects (UCRIHS) at Michigan State University. The form stated, and the trainer emphasized, that individual answers would only be seen by the trainer and that the overall results would be reported in an anonymous manner. In addition, the subjects were informed in the letter and by the trainer, that they would receive a copy of the results of this workshop. The consent forms were then signed, dated, and given to the trainer.

After signing this form the trainer passed out a copy of the rating scale. The subjects were asked to read the instructions and to make their ratings. After the ratings had been made, the subjects were instructed to turn the scale over and not to refer back to it during the remainder of the workshop.

The subjects were then provided with a copy of the motivation to rate accurately form. The trainer defined rating accuracy, reviewed the instructions with the subjects, asked them to carefully read each item, and answered any questions they had concerning the wording of items or the benchmark descriptions attached to each scale point. After the subjects had filled out this form, they were instructed to turn the form over and not to refer to it

during the remainder of the workshop.

The rating scale was completed prior to the motivation to rate accurately scale for two reasons. First, the hypotheses concerning the motivation to rate accurately assumed that the subjects had used one of the rating scales. Second, the motivation to rate accurately scale may have created an unwanted treatment effect. Those subjects that felt more motivated as a result of filling out the motivation to rate accurately scale may have made more accurate ratings than they would have if they had not completed this scale. In order to eliminate these alternative explanations to the data, the subjects filled out the rating scale before the motivation to rate accurately scale, and in both cases were instructed to turn over these scales and not to refer back to them once they had been completed.

After completing these two scales, two more forms were passed out to the subjects. The first one asked a series of demographic questions about the subjects. The second one was a reaction questionnaire where subjects indicated their reactions to the workshop. Both of these forms were described in a previous section of this chapter. Again, the trainer reviewed each form with the subjects and answered any questions.

After all four forms had been completed, the subjects were asked to clip them together. They were then collected by the trainer, the trainer promised to provide them with a copy and explanation of the results, and the workshop was concluded.

Analysis

Internal consistency estimates of reliability were assessed using Cronbach's alpha. The hypotheses were tested using a two-way analysis of variance (ANOVA). Because cell sample sizes were not equal nor proportional to one another, the means were not weighted by sample size (Keppel, 1973). When the F test was significant for a main effect, planned comparisons were used to identify significant differences between the means (Keppel, 1973). Statistical significance was assessed for the planned comparisons using the t statistic (Hays, 1973) and effect sizes were determined using omega - squared, ω^2 (Hays, 1963).

CHAPTER 5

Results

In this chapter the results for the tests of the hypotheses are presented. The analysis of variance, effect size, reliability, and correlational results are reported. A description of the support or lack of support for each hypothesis is presented. In the next chapter, the results are discussed.

Cronbach's alpha is reported in Table 1 for each of the scales used in this study. It can be seen that, with the exception of the instrumentality and expectancy scales, the reliability of these scales is adequate with coefficients ranging from .65 to .81.

Table 1. Scale Reliabilities

Scale	Alpha Coefficient
Behavior scale	.80
Trait scale	.81
Motivation to rate	.65
Expectancy	.56
Instrumentality	.41
Reactions	.65

The means, standard deviations, and intercorrelations are presented for all of the interval-level variables and all of the subjects in Table 2. The results in this table have a number of implications for subsequent data analyses. First, the correlation between the two dependent variables, Accuracy and Motivation to rate, is low and nonsignificant ($r=.03$, n.s.). Hence, a separate analysis of variance (ANOVA) was conducted for each dependent variable rather than combining the two variables and conducting a multivariate analysis of variance (MANOVA). Second, the correlations between five demographic variables (employees supervised, employees rated, supervisory experience, age, and education) and the dependent variables (performance rating accuracy and motivation to rate accurately) were all low and nonsignificant. Therefore, these five variables were not treated as covariates in subsequent analyses. Finally, the correlation between reactions and motivation to rate was moderate and significant ($r=.33$, $p<.001$). Consequently, the reactions measure was included as a covariate in the analysis of the motivation to rate accurately.

Sex, geographic location, and company training were dropped as covariates in the analyses of the dependent measures. A series of T-tests, shown in Tables 3, 4, and 5 revealed that there were no significant differences, for either rating accuracy or motivation to rate, between males and females, corporate and division employees, and employees trained and not trained in the company's performance appraisal system.

Table 2. Means, Standard Deviations, and Intercorrelations for Interval-Level Variables and All Subjects

Variable	$\bar{x}$	SD	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1. Behavior scale	1.66	.59												
2. Trait scale	3.12	.49	86											
3. Motivate to rate	4.17	.39	10	-02										
4. Expectancy	4.01	.44	08	04	90***									
5. Instrumentality	4.42	.49	08	-12	76***	40***								
6. Reactions	3.62	.66	15	-02	33***	36***	16							
7. Accuracy[a]	.95	.24	38***	14	03	02	03	-04						
8. Overall rating	2.56	.99	59***	66***	12	09	11	08	-05					
9. Employees supervised	20.09	31.36	00	-14	-04	-03	-02	-03	-03	-08				
10. Employees rated	7.51	5.60	20	02	-11	-06	-14	05	04	11	11			
11. Supervisory experience	6.19	6.34	11	06	-04	04	-15	-02	-03	67***	35***	04		
12. Education	5.25	1.25	06	04	-04	-06	01	-05	17	-03	-12	-04	-29**	
13. Age	40.74	9.66	01	16	-12	-07	-14	02	-08	05	19	14	67***	-07

[a]The signs of the correlations have been reversed
*p .05.
**p .01.
***p .001.

Table 3. Difference of Means Tests for Performance Rating Accuracy And Motivation to Rate Accurately By Sex

	Sex								
	Male			Female					
Variable	n	$\bar{x}$	SD	n	$\bar{x}$	SD	df	T	p <
Rating accuracy	66	.95	.25	19	.95	.20	83	.04	.97
Motivation to rate	66	4.16	.37	19	4.15	.45	83	.09	.93

Table 4. Difference of Means Tests for Performance Rating Accuracy And Motivation to Rate Accurately By Geographic Location

	Location								
	Corporate			Division					
Variable	n	$\bar{x}$	SD	n	$\bar{x}$	SD	df	T	p <
Rating accuracy	45	4.12	.40	41	4.22	.36	84	-1.28	.21
Motivation to rate	45	.95	.23	41	.95	.25	84	.17	.86

Table 5. Difference of Means Tests for Performance Rating Accuracy And Motivation to Rate Accurately By Company Training

	Company Training								
	Received			Not Received					
Variable	n	$\bar{x}$	SD	n	$\bar{x}$	SD	df	T	p <
Rating accuracy	64	4.14	.39	21	4.21	.36	83	-.68	.50
Motivation to rate	64	.95	.25	21	.97	.18	83	-.34	.74

The means and standard deviations for the reactions measure are presented by experimental condition in Table 6. As can be seen in Table 7, the effects of rater training, rating format, and their interaction were nonsignificant.

Table 6. Reaction Means and Standard Deviations By Experimental Condition

Condition	n	$\bar{x}$	SD
Trait Scale			
Error training	10	3.83	.76
Observation training	11	3.70	.75
Control group	19	3.42	.75
Behavior Scale			
Error training	12	3.56	.66
Observation training	18	3.67	.50
Control group	17	3.65	.65

Table 7. Analysis of Variance Results For Reactions

Source	df	MS	F	ω^2
Rating Format (F)	1	.00	.01	.00
Rater training (T)	2	.24	.53	.00
F x T	2	.44	.98	.00
Subj. w. groups	81	.45		

Rating Accuracy

The first four hypotheses treated performance rating accuracy as the dependent variable. In Table 8 the means and standard deviations for this variable are shown by experimental condition.

Table 8. Rating Accuracy Means and Standard Deviations by Experimental Condition

Condition	n	$\bar{x}$	SD
Trait Scale			
Error training	10	.76	.16
Observation training	11	.82	.17
Control group	19	.87	.21
Behavior Scale			
Error training	12	1.09	.25
Observation training	16	1.02	.22
Control group	17	1.02	.21

[a]The lower the score the more accurate the rating.

The first hypothesis stated that the use of a frequency of behavior scale would produce more accurate ratings than would a trait rating scale. As shown in Table 9 this rating scale effect was significant, $F(1,79)=24.22$, $p<.001$, and accounted for 20 percent of the rating accuracy variance ($\omega^2=.20$). However, the hypothesis was not supported as the means were opposite to the predicted direction (i.e. traits were rated more accurately).

Table 9. Analysis of Variance Results For Performance Rating Accuracy

Source	df	MS	F	ω^2
Rating Format (F)	1	.94	24.22*	.20
Rater training (T)	2	.01	.15	.00
F x T	2	.06	1.55	.00
Subj. w. groups	79	.04		

*p<.001.

The second and third hypothesis also failed to receive support. It was predicted that rater error training and observation training would be more accurate than the control group and in turn, that observational training would be more accurate than rater error training. The main effect for Rater training was nonsignificant, F(2, 79)=.15, n.s. The fourth hypothesis, that ratings will be more accurate when the rating format and rater training are consistent with one another, was also not confirmed as the Rating format x Rater training interaction was nonsignificant, F(2, 79)=1.55, n.s.

Motivation

The second set of hypotheses treated the motivation to rate accurately as the dependent variable. From Table 2 it can be seen that the two motivation to rate accurately subscales, instrumentality and expectancy, are significantly correlated with the motivation to rate accurately scale. In addition, as expectancy theory would predict given the

training intervention in this experiment, the correlation between expectancy and the motivation to rate accurately was larger than the correlation between instrumentality and the motivation to rate accurately, and was also larger than the correlation between expectancy and instrumentality.

In Table 10 the means and standard deviations for the motivation to rate accurately are shown by experimental condition. The ANOVA and analysis of covariance (ANOCOVA) results are presented in Tables 11 and 12. Again, there was no support for any of the four hypotheses treating motivation to rate accurately as the dependent variable.

The fifth hypothesis stated that the motivation to rate accurately would be greater when a frequency of behavior rather than a trait rating scale was used. This hypothesis

Table 10. Motivation to Rate Accurately Means And Standard Deviations By Experimental Condition

Condition	n	$\bar{x}$	SD
Trait Scale			
Error training	10	4.23	.44
Observation training	11	4.24	.39
Control group	19	4.19	.32
Behavior Scale			
Error training	12	4.12	.38
Observation training	18	4.18	.40
Control group	17	4.11	.46

Table 11. Analysis of Variance Results for Motivation to Rate Accurately

Source	df	MS	F	ω^2
Rating Format (F)	1	.14	.88	.00
Rater training (T)	2	.03	.19	.00
F x T	2	.44	.98	.00
Subj. w. groups	81	.16		

Table 12. Analysis of Covariance Results For Motivation to Rate Accurately

Source	MS	d.f.	F
Reactions	1.42	1	10.09*
Rating format (F)	.15	1	1.05
Rater training (T)	.01	2	.09
F x T	.01	2	.07
Residual	.14	80	

*p<.002

was not supported as indicated by the nonsignificant main effect for Rating format, F(1,81)=.88, n.s. Even when the results were adjusted for the reactions covariate this main effect was nonsignificant, F(1,80)=1.05, n.s.

It was predicted in the sixth and seventh hypotheses that the motivation to rate accurately would be greater for

the rater error training and observational training conditions than for the control group and that the motivation to rate accurately would be greater for observational training than for rater error training. These two hypotheses were not confirmed as the main effect for Rater training was nonsignificant for the ANOVA, $F(2,81)=.19$, n.s. and ANOCOVA, $F(2,80)=.09$, n.s., analyses.

The Rating format x Rater training interaction was nonsignificant when the data were analyzed using ANOVA, $F(2,81)=.02$, n.s., and ANOCOVA, $F(2,80)=.07$ n.s. Thus the eighth hypothesis, that the motivation to rate accurately will be greater when the rating format and the rater training are consistent with one another, was not supported.

Motivation and Rating Accuracy

The final hypothesis suggested that there would be a positive correlation between the motivation to rate accurately and performance rating accuracy. From Table 1 it can be seen that the correlation was positive, but very small in magnitude and at a nonsignificant level ($r=.03$, n.s.).

Summary

In summary, none of the eight hypotheses were supported by the tests of these data. The null hypothesis could only be rejected for the first hypotheses and it was in a direction opposite of the direction predicted. These results and their implications will be discussed in the next chapter.

CHAPTER 6

DISCUSSION

In this final chapter a discussion is presented concerning the lack of support for the hypotheses, the theoretical and applied implications of this research, the limitations associated with this study, and the directions that future research in this area might take. The chapter is organized by the hypotheses associated with each dependent variable and ends with a set of conclusions.

Rating Accuracy

The main effect for rating format was significant when performance rating accuracy was treated as the dependent variable. However, the direction of this relationship was opposite to the direction specified in hypothesis one. Traits rather than behaviors were rated more accurately. There are two potential sets of explanations for this finding. First, it may be the case that raters process performance information along trait-like dimensions. As schema theories suggest, raters' have preset categories to guide the observation, storage, and retrieval of stimuli (Alba & Hasher, 1983). These schema are usually global, trait-like dimensions and are often automatically used (Feldman, 1981). Consequently, it is not surprising that a

trait rating scale is more accurately rated as it more closely approximates the cognitive processes of the rater.

An equally likely set of explanations for this finding center around some limitations associated with this study. First, the frequency of behavior scale was extremely difficult to use, perhaps more demanding than what is actually required in field settings. Rarely are supervisors called upon to report the exact number of times a subordinate exhibits various behaviors. Second, the videotape observed by the subjects was lengthy and showed a large number of critical behaviors exhibited by the supervisor. To the extent that the subjects failed to pay strict attention to the videotape, and comments made by some of the subjects to the author suggested that they found the videotape to be uninteresting, it would be extremely difficult to keep track of the frequency of critical behaviors. Even if strict attention was given to the videotape, the performance was somewhat unrealistic as a large number of critical incidents were displayed in a compressed period of time. Finally, the videotape depcited a simulated set of work activities. This may have prevented a transfer of training from the workshop to the rating task.

The subjects using the trait rating scale did not have to pay attention to and memorize the frequency of critical behaviors and therefore, they may have been more accurate. These explanations are more consistent with the 'traditional' view of cognitive processing which suggests that the greater the demands on the memory of the rater, the less

accurate the rating (Heneman & Wexley, 1983).

The finding that traits are rated more accurately than behaviors presents some interesting implications. They must be tempered, however, by the methodological limitations just noted. In a very general sense this result supports Kavanaugh's (1971) contention that traits should not be automatically discounted as the content to be used in a performance appraisal system. He found little evidence to substantiate the claim that behaviors are superior to traits in terms of reliability and validity. Accuracy is also an important criterion in the evaluation of performance appraisal systems (Baird, 1982) and for the present sample, traits were rated more accurately than behaviors.

There are, however, a number of additional criteria that must be considered when evaluating performance appraisal systems. In particular, Feild and Holley (1982) have provided evidence which suggests that traits are not defensible in a court of law, Brumback (1972) has argued that traits have little relevance, and Patten (1982) and Latham and Wexley (1981) have reviewed evidence which suggests that traits are poor for employee feedback and development purposes, and for user acceptance. Consequently, an endorsement of trait rating scales is not warranted from this study. Furthermore, the accuracy of other methods of performance appraisal (e.g., MBO and employee comparisons) relative to traits have not been investigated.

At a more theoretical level, the finding that traits are more accurate than ratings has implications for future

research. It suggests that schema theories may be useful in coming to a better understanding of the rating process. Raters may automatically process ratings along trait dimensions. Before a firm conclusion like this can be drawn, however, a more direct test of this hypothesis needs to be made. If this hypothesis is confirmed by future research, and to the extent that performance rating systems other than traits are to be used, then more attention must be given to devising methods to shift the raters schema from traits to behaviors or results.

The second and third hypotheses predicted that ratings would be more accurate for those subjects receiving observation training than for those receiving rater error training and in turn, the accuracy of ratings in both of these conditions would be greater than the control group. These two hypotheses were not supported as indicated by the nonsignificant main effect for rater training. This lack of support may be due to several factors. First, the subjects were much older and more experienced than the trainer, and there were no rewards or sanctions associated with attendance at the seminar. Consequently, there may have been a limited amount of learning for the experimental and control groups.

Second, the rater error training program was of short duration and because of this time limitation, did not include an exercise on eliminating the "contrast" effect that normally is included in the program developed by Latham et al. (1975). Evidence reviewed by Spool (1978) suggests

that the training program must be of long duration for it to increase accuracy; and Latham and Wexley (1981) argue for the pervasiveness of contrast effects in appraisal judgments.

Third, the content of the rater training may need to focus on the actual rating instrument and categories to be used, rather than only focusing on judgment errors or what to observe. Pulakos (1983) found that training which focuses on the rating scale produced more accurate ratings than did rater error training. In this type of training emphasis is placed upon the transfer of an element (the rating scale) rather than principles (e.g., eliminating common rater errors) of the rating task (Royer, 1979). Perhaps this method is more effective in altering the trait oriented schema used by raters.

Fourth, these data indicate that the training was not effective for a work sample test (i.e., the videotaped performance of a supervisor), but do not speak to the issue of whether the training was transferred back to the ratings made on the job. It was not possible to gather this data because of the need to have a 'true' score with which to calculate accuracy.

The transfer of training is, however, an important issue for rater training and future researchers may wish to examine various methods to increase the transfer of learning. A number of leads have been offered in the literature including goal setting and positive reinforcement (Anderson & Wexley, 1983; Wexley & Nemeroff, 1975), relapse prevention training where managers learn to identify and

cope with situations that may eliminate the newly learned behaviors (Marx, 1982), and making the stimulus material in the training similar to the stimuli faced on the job (Wexley, in press; Wexley & Latham, 1981). In addition, Baumgartel, Sullivan, and Dunn (1978) have identified factors in the climate of the organization (e.g., growth orientation) that facilitate the transfer process. Finally, transfer may be facilitated by monetary or nonmonetary rewards (i.e., holding the person accountable for the transfer of training).

The third hypothesis predicted that accuracy would be greater when the rater training program and rating format were consistent with one another. This hypothesis was not confirmed as the Rating Format X Rater Training interaction was nonsignificant. It may be the case here that the subjects continued to rely on trait oriented schema after all types of training. This conclusion seems reasonable given the short duration of the training programs. In addition, this result may be due to the possibility that judgment and observation in the rating task are highly intercorrelated with one another. Thus, training on judgment errors is important for an observation based rating task and training on observational skills is important for a rating task requiring judgment.

These explanations have several implications. First, cognitive processing theories may be helpful in coming to an understanding of the rating process, but may have less utility in the design of a program to increase accuracy.

Second, because the judgment stage and observation stage of cognition in the rating task are interrelated, developers of rater training programs may wish to emphasize both observation and judgment skills. Finally, more emphasis may need to be placed in rater training programs on getting raters to shift from trait oriented schema to the categories used on the organizations rating scale. Alternatively, the dimensions of performance on the rating scale may need to be labeled using trait definitions. This is a common practice with another area of performance evaluation--the assessment center (Bray, Campbell, & Grant, 1974).

Motivation

The fifth though seventh hypotheses were concerned with the motivation to rate accurately. In particular, it was predicted that:

- Motivation to rate accurately will be less for raters using a trait rating scale than for those using a frequency of behavior scale.
- Raters given training that provides practice and feedback on the accuracy of their ratings will be more motivated to rate accurately than raters not receiving this practice and feedback.
- Raters given observational training will be more motivated to rate accurately than those given rater error training.
- The motivation to rate accurately will be greater when the rater training and rating format are con-

sistent with one another.

The results did not support these hypotheses. Both main effects, rater training and rating format, and the rater training x rating format interaction were nonsignificant. One likely reason for this set of findings is that rater's are highly motivated to rate accurately. In the present study, this appeared to be true regardless of the type of rating format used or type of rater training given to the subjects. On a 5 point Likert-type scale, with a 1 indicating low motivation and a 5 representing high motivation, the mean scores for all six conditions ranged from 4.11 to 4.24 with the standard deviations ranging from .32 to .46. Given the importance of performance ratings to the employer and employee, it may be the case that very little prompting is needed to get supervisors to work hard at making accurate ratings.

Another set of reasons for these results have to do with the experiment and the rating scale. First, the experimenter may have introduced unwanted demand effects. The subjects may have given high ratings because they felt that is what the experimenter and/or organization wanted from them. This explanation seems doubtful, however, as the subjects' responses were kept anonymous. Second, the word "accuracy" used in items on the scale may have been misinterpreted by the subjects as meaning how clearly the supervisor could express his opinions about the employee to be rated. It could also be argued that the wording of the items on the scale better reflected the subjects' perceived

skills at making accurate ratings than their motivation to rate accurately. Finally, the items were all worded in a highly positive manner and the subjects may have given what they perceived to be the socially desirable response.

These findings suggest that raters want to make accurate ratings regardless of the type of rating format or rater training received. Before a firm conclusion can be reached, however, more scale development is necessary. In particular the items need to be reworded and the definition of the construct may need to be broadened from the motivation to rate accurately to the motivation to rate. This broader definition might encompass all stages of the rating process including feedback (i.e., the rater may be motivated to make accurate ratings, but not be motivated to feed the ratings back to the employee). Finally, this construct needs to be validated in a field setting with a minimum of demand effects.

Motivation and Rating Accuracy

The final hypothesis stated that there would be a positive correlation between performance rating accuracy and the motivation to rate accurately. The correlation in this study was positive, but small in magnitude and nonsignificant. Given the high motivation to rate accurately values and the low variance, this result is not surprising. It does suggest, however, that the need to consider motivation in rating accuracy models is less important than the need to look at the skills and abilities of the rater to make

accurate ratings.

Conclusions

The results of this study suggest that raters cognitively process performance information using trait oriented schema. Consequently their ratings are likely to be more accurate when a trait rating scale is used rather than a frequency of behavior scale. If a frequency of behavior scale is to be used at the same level of accuracy it would appear that more emphasis needs to be placed in rater training on shifting the schema used by rater's from traits to behaviors.

The findings in this study also suggest that raters are highly motivated to make accurate ratings. Researchers and organizations interested in the prediction, understanding, and control of performance rating accuracy may, therefore, wish to focus more attention on the skills and abilities of the rater to make accurate ratings rather than the motivation of the rater to make accurate ratings. Both sets of conclusions must, however, be treated as being tentative given the methodological limitations to this study previously described.

Organizations that would like to increase the accuracy of performance ratings can address this issue in two ways on the basis of these conclusions. First, emphasis should be placed on developing the skill levels rather than the motivational levels of raters as they engage in the rating task. Second, training programs designed to increase the

skill levels of raters should focus on the dimensions of performance that are to be rated. If behaviors rather than traits are to be rated, then a program of long duration may be needed to shift the schema used by raters from traits to behaviors.

Finally, this dissertation points to the limitations associated with studies of performance rating accuracy that utilize student samples. It appears that the results from studies using student samples may not generalize to working supervisors. It has been demonstrated, for example, that students make more accurate ratings using behaviors rather than traits (e.g., Fay and Latham, 1982). In the present study, however, it was shown that supervisors rate traits more accurately than behaviors. It may be the case that students are trained, through their experiences in the classroom, to process discrete units of information whereas supervisors, in a very busy work environment, may rely upon more general, trait-like schema to process information. Consequently, more traditional theories of cognitive processing may be relevant for students while schema theories may be more applicable for supervisors. More emphasis should be placed upon obtaining industrial samples in future performance rating accuracy research.

APPENDICES

APPENDIX A

Frequency of Behavior Scale

For each of the statements listed below, circle the number that indicates the number of times you saw Jim Bogi, production supervisor, doing the behavior described.

1. Insisted that subordinates build the product in a certain way.

   0 1 2 3 4

2. Brought problems he had working with a subordinate to the subordinate's attention.

   0 1 2 3 4

3. Made sure that subordinates knew what to do while he was away.

   0 1 2 3 4

4. Used good suggestions brought up by subordinates.

   0 1 2 3 4

5. Pitched in and helped subordinates with their work.

   0 1 2 3 4

6. Refused to listen to a subordinate's request.

   0 1 2 3 4

7. Instructed subordinates on the proper quality of the product.

   0 1 2 3 4

8. Listened patiently to a subordinate's gripes.

   0 1 2 3 4

9. Emphasized the need for faster production.

   0 1 2 3 4

Frequency of Behavior Scale (Continued)

10. Solicited subordinate's ideas and opinions on what parts to purchase and what products to sell.

0 1 2 3 4

11. Praised subordinates for good suggestions.

0 1 2 3 4

12. Kept careful track of the profit margin.

0 1 2 3 4

13. Made his supervisory duties clear to subordinates.

0 1 2 3 4

14. Planned in advance the products to be built.

0 1 2 3 4

15. Offered suggestions on the best method to build the product.

0 1 2 3 4

16. Refused to listen to subordinate's suggestions.

0 1 2 3 4

17. Recognized his own weaknesses and asked a subordinate for his assistance in these matters.

0 1 2 3 4

18. Constructively criticized a subordinate when the subordinate made an error.

0 1 2 3 4

19. Gave subordinates work assignments.

0 1 2 3 4

20. Guided subordinates on the products to be manufactured.

0 1 2 3 4

Frequency of Behavior Scale (Continued)

21. Solicited subordinates opinions and ideas on what product to build.

0 1 2 3 4

APPENDIX B

Trait Rating Scale

Indicate the degree to which you agree or disagree with the following statements concerning the performance of Jim Bogi, production supervisor. Circle one number for each statement.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1. Is enthusiastic about job.	1	2	3	4	5
2. Has poor emotional balance.	1	2	3	4	5
3. Stalls on job.	1	2	3	4	5
4. Cannot be trusted.	1	2	3	4	5
5. Upsets morale.	1	2	3	4	5
6. Is proud of work.	1	2	3	4	5
7. Often forgets.	1	2	3	4	5
8. Is active and energetic.	1	2	3	4	5
9. Seldom sticks to business.	1	2	3	4	5
10. Is self controlled.	1	2	3	4	5
11. Is always complaining.	1	2	3	4	5
12. Is hard to get along with.	1	2	3	4	5
13. Is lazy.	1	2	3	4	5
14. Loses temper easily.	1	2	3	4	5
15. Has common sense.	1	2	3	4	5

Trait Rating Scale (Continued)

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
16. Worries occasionally.	1	2	3	4	5
17. Lacks initiative.	1	2	3	4	5
18. Sometimes does not fit into group.	1	2	3	4	5
19. Is pessimistic.	1	2	3	4	5
20. Is slow to adjust.	1	2	3	4	5

APPENDIX C

Overall Rating

Overall, how would you rate Jim Bogi's performance? Circle one number.

Poor	Below Average	Average	Above Average	Excellent
1	2	3	4	5

APPENDIX D

Motivation to Rate Accurately

Indicate the degree to which you disagree with each of the following statements. Circle one number for each statement.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1. I am able to make accurate performance ratings.	1	2	3	4	5
2. It is important to make accurate performance ratings.	1	2	3	4	5
3. The harder I work at it the more accurate my performance ratings will be.	1	2	3	4	5
4. I know how to make more accurate performance ratings.	1	2	3	4	5
5. I am concerned about the accuracy of performance ratings.	1	2	3	4	5
6. It is possible for me to make my performance ratings more accurate.	1	2	3	4	5
7. I am interested in making more accurate performance ratings.	1	2	3	4	5
8. I am confident that I can make more accurate performance ratings.	1	2	3	4	5

APPENDIX E

Reactions

Content

Indicate your reactions to the information that was presented to you in this workshop. Circle one number and write in your comments.

Poor	Below Average	Average	Above Average	Excellent
1	2	3	4	5

Comments:

Process

Indicate your reactions to how the information in this workshop was presented to you. Circle one number and write in your comments.

Poor	Below Average	Average	Above Average	Excellent
1	2	3	4	5

Comments:

Reactions (Continued)

Recommendation

Indicate the degree to which you agree or disagree with the following statement. I would recommend that other supervisors attend this workshop. Circle one number and write in your comments.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	2	3	4	5

Comments:

APPENDIX F

Demographics

Please fill in the blank for each of the following statements.

1. I am ____years old.

2. The title of my position is _____.

3. I supervise a total of _____ employees.

4. I make performance ratings for a total of _____ employees.

5. I have been a supervisor for _____ years.

6. I work in the _____ division.

7. I work in the _____ department.

Please circle one letter for each of the following statements.

1. I am:

   (a) Male
   (b) Female

2. I have been trained on how to use _____ the performance review system.

   (a) True
   (b) False

3. My education level is:

   (a) Grade school
   (b) Some high school, no diploma
   (c) High school diploma
   (d) Some college, no degree
   (e) College degree
   (f) Some graduate school, no degree
   (g) Graduate school degree

REFERENCES

REFERENCES

Alba, J. W., & Hasher, L. Is memory schematic? Psychological Bulletin, 1983, 93, 203-231.

Anderson, J. G., & Wexley, K. N. Application-based management development. Personnel Administrator, 1983, 28, 39-44.

Atkin, R. S., & Conlon, E. J. Behaviorally anchored rating scales: Some theoretical issues. Academy of Management Review, 1978, 3, 119-138.

Baird, L. S. Why worry about accurate measures? In L. S. Baird, R. W. Beatty, & C. E. Schneier (Ed.'s). The performance appraisal sourcebook. Amherst, MA: Human Resource Development Press, 1982, 12-16.

Bartlett, C. J. What's the difference between valid and invalid halo? Forced-choice measurement without forcing a choice. Journal of Applied Psychology, 1983, 68, 218-226.

Baumgartel, H., Sullivan, G. J., Dunn, L.E. How organizational climate and personality affect the pay-off from advanced management training sessions. Kansas Business Review, 1978, 5, 1-10.

Bayroff, A. G., Haggerty, H. R., & Rundquist, E. A. Validity of ratings as related to rating techniques and conditions. Personnel Psychology, 1954, 7, 93-113.

Berman, J. S., & Kenny, D. A. Correlational bias in observer ratings. Journal of Personality and Social Psychology, 1976, 34, 263-273.

Bernardin, H. J., & Beatty, R. W. Performance appraisal: Assessing human behavior at work. Boston: Kent, 1984.

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

Bernardin, H. J., & Cardy, R. L. Appraisal accuracy: The ability and motivation to remember the past. Public Personnel Management Journal, 1982, 11, 352-357.

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

Boice, R. Observational skills. Psychological Bulletin, 1983, 93, 3-29.

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

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

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

Borman, W. C. Individual differences correlates of accuracy in evaluating others' performance. Applied Psychological Measurement, 1979a, 3, 103-115.

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

Borman, W. C. Implications of personality theory and research for the rating of work performance in organizations. In F. Landy, S. Zedeck, and J. Cleveland (Ed.'s), Performance Measurement and Theory, Hillsdale, NJ: Lawrence Erlbaum, 1983, 127-165.

Bray, D. W., Campbell, R. J., and Grant, D. L. Formative years in business: A long-term AT&T study of managerial lives. New York: Wiley, 1974.

Brumback, G. B. A reply to Kavanagh's: "The content issue in performance appraisal: A review." Personnel Psychology, 1972, 25, 567-572.

Bureau of National Affairs. (1983). Performance Appraisal Programs. (Personnel Policies Forum Survey No. 135). Washington, D.C.: The Bureau of National Affairs, Inc.

Campbell, D. T. The mutual methodological relevance of anthropology and psychology. In F. L. K. Hsu (Ed.), Psychological Anthropology. Homewood, IL: Dorsey, 1961, 333-352.

Carroll, S.J., & Schneier, C.E. Performance appraisal and review systems. Glenview, IL: Scott, Foresman, and Company, 1982.

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

Cronbach, L.J. Processes affecting scores on "understanding of others" and "assumed similarity." Psychological Bulletin, 1955, 52, 177-193.

Cronbach, L.J., & Furby, L. How should we measure "change" - or should we? Psychological Bulletin, 1970, 74, 68-80.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. The dependability of behavior measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

DeCotiis, T. & Petit, A. The performance appraisal process: A model and some testable propositions. Academy of Management Review, 1978, 3, 635-646.

Downs, C.W., & Moscinsky, P. (1979). A survey of appraisal processes and training in large corporations. Paper presented at the 39th Annual Academy of Management Meetings, Atlanta, GA.

Eder, R.W., Keaveny, T.J., McGann, A.F., & Beatty, R.W. Evaluating faculty performance: An empirical investigation of factors affecting faculty ratings and student satisfaction using alternative rating forms. Proceedings of the 38th Annual Meeting of the Academy of Management, 1978, 13, 23-27.

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

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

Feild, H.S., & Holley, W. The relationship of performance appraisal system characteristics to verdicts in selected employment discrimination cases. Academy of Management Journal, 1982, 25, 392-406.

Firth, R. Elements of social organization, 3ed. Boston: Beacon Press, 1961.

Flanagan, J.C. A new approach to evaluating personnel. Personnel, 1949a, 10, 35-42.

Flanagan, J.C. Critical requirements: A new approach to employee evaluation. Personnel Psychology, 1949b, 2, 419-425.

Flanagan, J.C. Principles and procedures in evaluating performance. Personnel, 1952, 28, 373-386.

Flanagan, J.C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-358.

Flanagan, J.C., & Burns, R.K. The employee performance record. Harvard Business Review, 1957, 35, 95-102.

Gibson, J.J. The concept of the stimulus in psychology. American Psychologist, 1960, 15, 694-703.

Gordon, M.E. The effect of the correctness of the behavior observed on the accuracy of ratings. Organizational Behavior and Human Performance, 1970, 5, 366-377.

Gordon, M.E. An examination of the relationship between the accuracy and favorability of ratings. Journal of applied Psychology, 1972, 56, 49-53.

Hays, W.L. Statistics for social sciences, 2ed. New York: Holt, Rinehart, and Winston, 1963.

Hays, W.L. Statistics for psychologists, New York: Holt, Rinehart, and Winston, 1967.

Heneman, R.L. & Wexley, K.N. The effects of time delay in rating and amount of information observed on performance rating accuracy. Academy of Management Journal, 1983, 26, 677-686.

Ilgen, D.R. Gender issues in performance appraisal: A discussion of O'Leary and Hansen. In F. Landy, S. Zedeck, & J. Cleveland (Ed's). Performance theory and measurement. Hillsdale, NJ: Lawrence Erlbaum, 1983, 219-220.

Johnson, D.M. A systematic treatment of judgment. Psychological Bulletin, 1945, 42, 193-224.

Kavanagh, M.J. The content issue in performance appraisal: A review. Personnel Psychology, 1971, 24, 653-668.

Kelley, M. A contingency framework for evaluation. Academy of Management Review, 1978, 5, 428-438.

Keppel, G. Design and analysis. Englewood Cliffs, NJ: Prentice Hall, 1973.

Kohler, W. Gestalt psychology, New York: Liversight, 1956.

Kraiger, K. (1983, March). Cognitive processes in rating bias. Paper presented at the I/O-OB Graduate Student Conference, Chicago, IL.

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

Landy, F.J., & Farr, J.L. The measurement of work performance. New York: Academic Press, 1983.

Latham, G.P., Wexley, K.N., & Purcell, E.D. Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 1975, 60, 550-555.

Latham, G.P., & Wexley, K.N. Increasing productivity through performance appraisal. Reading, MA: Addison-Wesley, 1981.

Lawler, E.E. Pay and organizational effectiveness. New York: McGraw-Hill, 1971.

Lawlis, G.F., & Lu, E. Judgment of counseling process: Reliability, agreement, and error. Psychological Bullegin, 1972, 78, 17-20.

Locke, E.A., Shaw, K.N., Saari, L.M., & Latham, G.P. Goal setting and task performance: 1969-1980. Psychological Bulletin, 1981, 90, 125-152.

Lopez, F.M. Evaluating employee performance. Chicago: Public Personnel Association, 1968.

Maier, N.F., & Thurber, J.A. Accuracy of judgments of deception when an interview is watched, heard, and read. Personnel Psychology, 1968, 21, 23-30.

Marx, R.D. Relapse prevention for managerial training: A model for maintenance of behavior change. Academy of Management Review, 1982, 7, 433-441.

McGregor, D. An uneasy look at performance appraisal. Harvard Business Review, 1957, 35, 89-94.

Mitchell, T.R. Expectancy model of job satisfaction, occupational preference and effort: A theoretical, methodological, and empirical appraisal. Psychological Bulletin, 1974, 81, 1003-1077.

Mohrman, A.M., Jr., & Lawler, E.E., III. Motivation and performance appraisal behavior. In F. Landy, S. Zedeck, & J. Cleveland (Ed's) Performance measurement and theory, Hillsdale, N.J.: Lawrence Erlbaum, 1983, 173-189.

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

Murphy, K.R., Martin, C., & Garcia, M. Do behavioral observation scales measure observation? Journal of Applied Psychology, 1982, 67, 562-567.

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

Naylor, J.C. Some comments on the accuracy and the validity of a cue variable. Journal of Mathematical Psychology, 1967, 4, 154-161.

Osburn, H.G., Timmreck, C., & Bigby D. Effect of dimensional relevance on accuracy of simulated hiring decisions by employment interviewers. Journal of Applied Psychology, 1981, 66, 159-165.

Patten, T.H., Jr. A manager's guide to performance appraisal. New York: Free Press, 1982.

Pulakos, E.D., (1983). A comparison of two rater training programs: Error training versus accuracy training. Unpublished master's thesis, Michigan State University, East Lansing, MI.

Purcell, T.V. Observing people. Harvard Business Review, 1955, 33, 90-100.

Richards, J.M., Jr., & Cline, V.B. Accuracy components in person perception scores and the scoring system as an artifact in investigations of the generality of judging ability. Psychological Reports, 1963 12, 363-373.

Rogasa, D., Brandt, D., & Zimowski, M. A growth curve approach to the measurement of change. Psychological Bulletin, 1982, 92, 726-748.

Royer, J.M. Theories of the transfer of training. Educational Psychologist, 1979, 14, 53-69.

Rush, M.C., Phillips, J.S., & Lord, R.G. Effects of temporal delay in rating of leader behavior descriptions: A laboratory investigation. Journal of Applied Psychology, 1981, 66, 442-450.

Smith, M. Documenting employee performance. In L.S. Baird, R.W. Beatty, & C.E. Schneier (Ed.'s). The performance appraisal sourcebook. Amherst, MA Human Resource Development Press, 1982, 94-96.

Spool, M.D. Training programs for observers of behavior: A review. Personnel Psychology, 1978, 31, 853-888.

Taft, R.L. The ability to judge people. Psychological Bulletin, 1955, 52, 1-23.

Thorndike, R.L. Personnel Selection. New York: John Wiley & Sons, 1949.

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

Tinsley, H. A., & Weiss, D. J. Interrater reliability and agreement on subjective judgements. Journal of Counseling Psychology, 1975, 22, 358-376.

Uhrbrock, R.S. 2000 scaled items. Personnel Psychology, 1961, 14, 375-420.

Vroom, V.H. Work and motivation. New York: Wiley, 1964.

Wakeley, J.H. (1964). The effects of specific training on accuracy in judging others. Unpublished dissertation, Michigan State University, East Lansing, MI.

Warmke, D.L. Effects of accountability procedures upon the utility of peer ratings of present performance. (Doctoral dissertation, Ohio State University, 1979). Dissertation Abstracts International, 1980, 40, 4011-B. (University Microfilms No. 80-01, 853).

Weick, K.E. Systematic observational methods. In G. Lindzey & E. Aronson (Ed.'s). The handbook of social psychology. Reading, MA: Addison-Wesley, 1968, 357-451.

Weick, K.E. The social psychology of organizing. Reading, MA: Addison-Wesley, 1979.

Wexley, K.N. Personnel training. Annual Review of Psychology, 1984, in press.

Wexley, K.N. (1982, November). The performance appraisal interview. Paper presented at the Fourth John Hopkins University National Symposium on Educational Research, Washington, DC.

Wexley, K.N., & Jaffee, C.L. Evaluation of the telecoaching training method. Journal of Industrial Psychology, 1970, 5, 58-62.

Wexley, K.N., & Latham, G.P. Developing and training human resources in organizations. Glenview, IL: Scott, Fovesman, and Company, 1981.

Wexley, K.N., & Nemeroff, W.F. Effectiveness of positive reinforcement and goal setting as method of management development. Journal of Applied Psychology, 1975, 60, 446-450.

Wexley, K.N., Sanders, R.E., & Yuk, G.A. Training interviewers to eliminate contrast effects in employment interviews. Journal of Applied Psychology, 1973, 57, 233-236.

Wexley, K.N., & Youtz M.A., (1983) Rater values: Their effects on rating errors and rater accuracy. Unpublished manuscript.

Wherry, R.J., The control of bias in ratings: A theory of rating. PRB Report No. 922, Contract No. DA-49-083, OSA 69, Department of the Army, 1952.

Wherry, R.J., Sr. & Bartlett, C.J. The control of bias in ratings. Personnel Psychology, 1982, 35, 521-552.

Wiggins, J. Personality and prediction: Principles of personality assessment. Reading, MA: Addison-Wesley, 1973.

Wrightsman, L. Measurment of philosophies of human nature. Psychological Reports, 1964, 14, 743-751.

Zedeck, S., & Cascio, W.F. Performance appraisal decisions as a function of rater training and purpose of the appraisal. Journal of Applied Psychology, 1982, 67, 752-758.