This is to certify that the

thesis entitled

A Comparison of Two Rater Training Programs:
Error Training Versus Accuracy Training

presented by

Elaine Diane Pulakos

has been accepted towards fulfillment
of the requirements for

M. A. degree in Psychology

Neal Schmitt

Major professor

Date 3/18/83

A COMPARISON OF TWO RATER TRAINING PROGRAMS: ERROR TRAINING VERSUS ACCURACY TRAINING

By

Elaine Diane Pulakos

A THESIS

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

MASTER OF ARTS

Department of Psychology

1983

ABSTRACT

A COMPARISON OF TWO RATER TRAINING PROGRAMS: ERROR TRAINING VERSUS ACCURACY TRAINING

By

Elaine Diane Pulakos

The purpose of this research was to investigate the effects of Rater Error Training (RET) and Rater Accuracy Training (RAT) on two rating errors (halo and leniency) and accuracy of performance evaluations. Differences in program effectiveness for various job performance dimensions were also assessed. One hundred and eight subjects were randomly assigned to 1 of 4 cells defined by the training treatments (RET, RAT, RET and RAT, no training), and raters evaluated videotaped ratees. The results showed that RAT increased accuracy and decreased leniency, while RET decreased halo but had no effect on leniency or accuracy. The combination of RET and RAT yielded less accurate ratings than RAT alone. Finally, dimension x training interactions suggested that the effectiveness of training strategies can not be considered independent of the rating format. Implications and directions for future research are discussed.

ACKNOWLEDGMENTS

First and foremost, this project would not have been possible without my parents, who have provided me with a great deal of encouragement as well as the opportunity to pursue my goals. My sincere appreciation also goes to Neal Schmitt for his guidance and support on this particular project and in general. Next, I would like to thank my other committee members, Ken Wexley and John Wagner for their helpful suggestions. Last, but not least, my gratitude goes to a fourth but unofficial committee member, Arnon E. Reichers, for her moral support, abstract intellectual abilities, and for teaching me a lot about writing and organizing papers.

Finally, I would like to dedicate this thesis to my brother, George, who has been with me throughout this entire project. He has helped in more ways than can possibly be mentioned here and has made the frustrating times a little more tolerable.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The most widely used type of instrument for obtaining performance measures is the rating scale (Borman, 1979). Many of the personnel decisions made in organizations rely on the ability of supervisory ratings to discriminate good performers from poor performers. Furthermore, ratings are often the only criteria available for validating selection procedures, promoting employees, and selecting individuals for training programs. A major problem with performance ratings, however, is that they are inevitably contaminated by various rater errors which render them of questionable reliability, validity, and accuracy (Bernardin & Pence, 1980). Specifically, rater errors are faults in judgment that occur in a systematic manner when one individual evaluates another (Latham & Wexley, 1981). Some of the more commonly cited of these are halo, central tendency, leniency, and strictness (Guilford, 1954). The problems imposed by such errors have led many researchers to call for the development of rater or observer training programs to improve the quality of performance evaluations (e.g., DeCotiis & Petit, 1978; Dunnette & Borman, 1979).

Many of the rater training programs to date have been successful in reducing common rating errors such as halo and/or leniency, at least as they have been statistically measured (Bernardin, 1978; Bernardin & Walter, 1977; Borman, 1975; Ivancevich, 1979; Latham, Wexley, & Pursell,

1975). A common assumption among these researchers, however, is that reducing psychometric errors will also result in increasing performance rating accuracy (i.e., that error and accuracy negatively covary). Accuracy has been defined as the degree to which ratings are relevant to or correlated with true criterion scores (Dunnette & Borman, 1979). In his review of observer training programs, Spool (1978) concluded that studies assessing training effects indicate that "accuracy in observation can be improved by training raters to minimize rating errors" (pp. 866-867).

Unfortunately, the assumption that errors and accuracy negatively covary has for the most part been unaddressed empirically. This state of affairs is largely the result of error reduction strategies focusing on rating behavior while largely ignoring the issue of accuracy. Recent rating accuracy research, however, has raised questions regarding the prevailing error/accuracy covariation assumption (Berman & Kenny, 1977; Borman, 1975, 1979; Warmke, 1980). Specifically, the data from this research seem to suggest not only that rating accuracy is largely unaffected by training, but that there may even be a weak positive relationship between certain errors (e.g., halo) and accuracy (Cooper, 1981). These results not only run counter to a basic tenet of psychometric theory (i.e., error produces inaccuracy), but they raise serious questions regarding the utility of most rater training efforts to date.

Although rater training programs have differed with respect to some of their key components (e.g., level of trainee participation, feedback to participants, amount of practice time allowed), a common core to

virtually all training efforts has been a general concern for training aimed at changing rater response distributions (Bernardin & Pence, 1980). Landy and Farr (1980) and Wherry (Note 1) have proposed a tenable hypothesis for why this focus may have little effect on improving accuracy. Specifically, these authors have suggested that concern with psychometric error distributions alone merely facilitates the learning of new response sets. The programs may thus achieve lower mean ratings (i.e., less leniency) and lower scale intercorrelations (i.e., less halo) but perhaps lower levels of accuracy as well. This reasoning was based on the possibility that skewed ratings and high dimension intercorrelations may reflect reality (Schwab, Heneman, & DeCotiis, 1975). Based upon these arguments, it seems logical that increasing accuracy may require focusing trainee attention directly on accuracy issues rather than concentrating solely on rating errors.

The purpose of the present research was to assess differences in rater training as a function of the orientation of two rater training programs. Rater error training (RET) similar to that developed by Latham et al (1975) was compared to a type of rater accuracy training (RAT). Rather than training to reduce errors, per se, the focus of the accuracy training program was to familiarize raters with the instrument/rating scale and focus their attention to the specific behaviors they would be asked to evaluate. Drawing on literature from cognitive psychology (reviewed below), it was hypothesized that directing attention to appropriate aspects of the rating task itself and increasing rater familiarization with the instrument in a systematic way would have the desirable effect of increasing the accuracy of

observations.

In summary, the major thesis proposed here is that previous rater training efforts have erroneously focused rater attention to errors, rather than focusing attention to the observation of relevant ratee behaviors. It is further argued that this focus is largely a result of lack of attention to the cognitive processes involved in the rating task, and that this deficiency may be responsible for the error/accuracy covariation paradox. The next section presents a model which is used as a framework for a discussion of previous research.

The Rating Process

It has recently been suggested that without a better understanding of the cognitive processes involved in performance ratings and the variables influencing these, further gains in accuracy may be difficult to achieve (Cooper, 1981; Landy & Farr, 1980). While there are variations in cognitive process models of ratings, Feldman's (1981) model is both general and specific enough to be used as a basis for the present research. This model proposes that the cognitive processes involved in the rating task are a special case of a more generalized information processing model. Specifically, Feldman conceptualized the performance appraisal process as a combination of four interacting cognitive tasks (see Figure 1 for an illustration of this model). First, the rater must recognize and attend to relevant information concerning those who are being evaluated. Second, the information must be organized and stored for later access. New information must also be integrated with previously gathered data. The third step involves

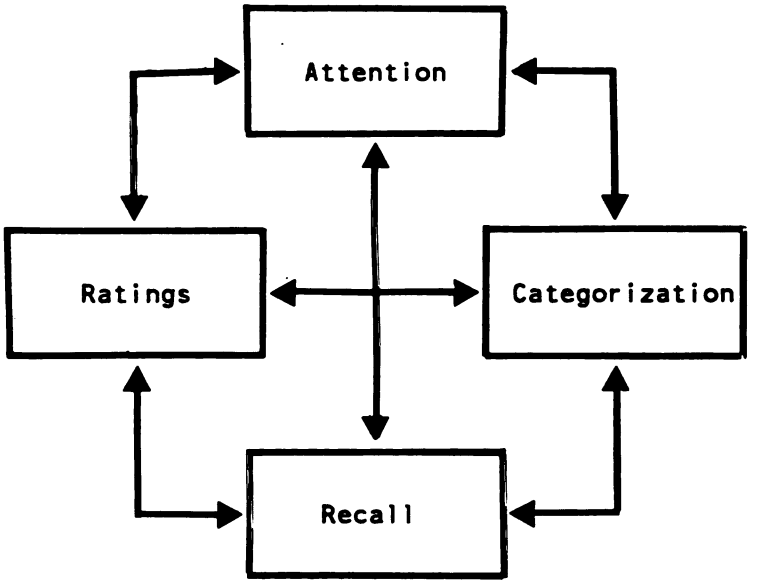


Figure 1. Feldman's Rating Process Model.

recalling relevant information in an organized fashion when judgments about performance are required. Finally, the rater must be able to integrate information into some kind of summary evaluation for most appraisal tasks.

While it may appear that these processes occur precisely in the order depicted in the model, Feldman (1981) cautions that they are interacting, dynamic, and cyclical. Thus, for example, previously formed categories may guide attention to certain stimuli while largely ignoring others, as well as forming the basis for subsequent categorizations and recall. What follows is a discussion of each of the components of Feldman's (1980) model as they relate to the present research. It is eventually argued that categories in a cognitive psychology sense are similar to the dimensions of a performance appraisal instrument, and that familiarity with these particular dimension structures cues raters to attend to relevant information. Relevant categorization and attention to appropriate cues (behaviors) presumably facilitate recall of pertinent information and should therefore be associated with increased rater accuracy.

The following section focuses on the previous research dealing with attention, categorization and recall processes. For the sake of greater clarity, each of these is discussed under a separate subsection. However, it should be remembered that the relationship among these categories is interacting and cyclical, resulting in some degree of overlap among them throughout the presentation. Following this review, a summarization and critique of previous observer training efforts is undertaken, along with a discussion of the present accuracy training

effort.

The Roles of Attention, Categorization, and Recall in Performance Appraisal

Attention

Individuals have a limited capacity to process the vast amount of information available at any given moment, and they must therefore be selective with respect to what is actually attended to on a conscious level (Glass, Holyoak, & Santa, 1979). There is a great deal of research, however, which indicates that the majority of everyday stimuli are automatically processed (Ableson, 1976; Langer, 1978; Nisbett & Wilson, 1977; Schneider & Shiffrin, 1977; Shank & Ableson, 1977; Shiffrin & Schneider, 1977). Race, sex, cues of dress, speech, height, attractiveness, etc. are all stimuli which can be automatically recorded (Feldman, 1981). For example, upon observing a woman, one does not typically ask, "is that a female and what difference does it make if she is?" One more generally recognizes sex automatically and unintentionally and thereafter reacts partially in terms of that classification.

Additional research indicates, however, that when cued, subjects can accurately recall those stimuli for which they have been prepared. For example, Averbach and Coriell (1961) conducted an experiment in which two rows of eight letters each were flashed in front of subjects for a tenth of a second. When subjects were subsequently asked to recall as many letters as possible, very few accurate recollections resulted. Subjects were then told to focus their attention to specific

positions on the screen (e.g., they were told to focus on the third letter in the third row). The vast majority of the participants were able to accurately recall those stimuli to which their attention had been directed. The results of similar research by Eriksen and Collins (1969) also showed the positive effects of directed attention in increasing recall accuracy.

In a related effort, Treisman and her colleagues (Treisman & Geffen, 1967; Treisman & Riley, 1969) also investigated the effects on recall of directing subjects' attention to specific cues. In a typical experiment, these reserachers simultaneously presented students with a list of digits to each ear, only one of which they were told would later have to be repeated. Occasionally, a letter was presented with the digits and students were instructed that when they heard the letter in either ear, they should tap their desks with a ruler. If the students had been equally aware of both the "attended" and the "unattended" ear, they should have detected the letter equally often in both ears. The results showed that subjects accurately detected about 80 percent of the letters presented to the attended ear and only about 23 percent of the letters presented to the unattended ear.

Finally, Lawrence (1971) used a tachistoscope to flash a list of words at a person one at a time (at a rate of 20 words per second). When a series of words was presented in this way, he found that subjects could accurately read very few, if any, of them. However, he additionally discovered that subjects could be cued in advance to read a particular word. In Lawrence's experiment, subjects were told that one word in the series would be in all capital letters, and they were to

focus on that word. The results showed that subjects were better able to identify the "target" word than when no cuing occurred. One conclusion drawn was that individuals' attention could be focused to a particular stimulus object rather than the entire modality (i.e., everything they saw). Further, because the "target" word was defined by a discriminating feature (i.e., capital letters), it enabled the participants to more effectively attend to it as well as to increase the accuracy in their recall of the word.

Taken as a whole, this research indicates that individuals can be cued to become consciously aware of particular stimulus objects in their sensory fields, and that this increased attention to specific features facilitates recall. A potential caution in interpreting these results in light of their relevance to the present research is that these experiments involved very simple attention and recall tasks (i.e., attending to letters/ words presented to subjects for short periods of time). The research proposed here attempts to build upon the theoretical conception discussed above by applying the notion of directed attention to more complex performance evaluation criteria. It is specifically argued that although previous rater training programs have cued raters to consciously attend to "relevant stimuli" (i.e., errors), this focus has been largely insufficient, especially because accuracy is the crucial criterion for judging performance rating quality and should therefore be the focus of rater training. A more complete rationale for this hypothesis is developed subsequent to the discussion of categorization and recall processes.

Categorization

Within a cognitive psychological framework, no discussion of attention, per se, is complete without a discussion of categorization. This is true not only because of the fact that these two processes are both components of Feldman's (1981) model, but also because the two concepts are intimately dependent upon each other in actuality. Bruner (1957, 1958) discusses this interdependence in his contention that conscious attention, hence perception, is the categorization of stimuli whereby individuals assign identity and meaning to an object. That is, individuals attend to and interpret their stimulus environment in terms of the cognitive categories most available to them. As such, hypotheses about category memberships follow from whatever categories the individual most typically uses to organize and make sense of the environment.

A category has been defined as a cognitive structure that partially consists of the representation of some defined stimulus domain. Categories can further be thought of as pyramid-like structures, organized with more general information at the top and more specific information nested within the more general groupings. The lowest level in the hierarchy consists of specific examples of category relevant objects/events. These organizational properties represent an individual's knowledge of the way in which the world is structured. When a stimulus configuration is encountered in the environment, it is matched to some category, and the ordering of the relations among the elements in the category are imposed on the elements of the stimulus configuration (Marcus, 1977; Minsky, 1975; Tesser, 1978). This process

of ordering and structuring the elements of the stimulus is important because it influences the subsequent recall of information and provides the basis for inferences and predictions (Taylor & Crocker, 1981).

While it is beyond the scope of the present proposal to review all of the relevant research involving categorization systems, it is worthwhile to note the central role that categories play in phenomena such as implicit personality theory (i.e., categorizations based on trait labels; Hastorf, Schneider, & Polefka, 1970; Lord, Binning, Rush, & Thomas, 1978) and stereotyping (categorizations based on cues such as race and sex; McArthur & Post, 1977; Taylor & Fiske, 1978). Further, Kelly's (1955) personal construct theory has delineated the sometimes profound individual differences that exist in individual category systems. For example, it has been shown that cultural factors (Triandis, 1964) and individual difference variables such as prejudice and cognitive complexity (Feldman & Hilterman, 1975) make different categories salient for different people. Additionally, situational factors (such as how often a category is used or how recently a category has been used; Wyer & Srull, 1980) affect which aspects of a given stimulus person or object will be used in categorization. Evidence supporting the notion that recall is dependent upon the category system employed by the perceiver is discussed in the following section.

Recall

When confronted with a stimulus configuration (e.g., person, object, or situation), one could conceivably recall any of a variety of stimulus attributes. Information is easier to recall, however, if it is

structured in some meaningful way. Further, there is evidence that people structure their observations so as to facilitate recall (Bousfield, 1953). Because categorizations provide a means of structuring and organizing what is observed, it has been suggested that either imposing a category system on stimulus configurations or encountering a stimulus configuration that is a good match to already established categories increases the recall of category relevant information.

This contention has been given empirical support by a number of research efforts. For example, Taylor, Livingston, & Crocker (1982) presented graduate students in different departments with an academic folder of a hypothetical student. Subjects were later tested on recall: English students recalled more English relevant material (e.g., English courses, languages, and writing skills), while psychology graduate students recalled more psychology relevant information (e.g., research experience, psychology courses, and math background), even though the experimental task did not require selective use of this material. Thus, the availability of previously existing category systems seemed to influence recall of certain types of information consistent with the categories already in use by the individual.

In a study on occupational stereotypes (Cohen, 1977), subjects observed a videotape of a woman performing some daily (non-work) activities, having been told either that she was a waitress or a librarian. In a free recall task, subjects recalled stereotype consistent information more accurately than irrelevant or inconsistent information. Other research has similarly demonstrated the effects of

imposed categorization systems for improving recall of category relevant information (Picek, Sherman, & Shiffrin, 1975; Potts, 1972; Sulin & Dooling, 1974; Woll & Yopp, 1978).

Recall of events and episodes has also been shown to be selectively improved by the imposition of category systems from external sources. For example,Zandy and Gerard (1974) had subjects observe a videotape of two people poking around an apartment. Some subjects were told the people were anticipating a drug bust and were looking for their dope so they could remove it. Others were told that the two were planning to rob the apartment, while a third group was told the two were waiting for a friend and had become stir crazy. The results showed that subjects remembered more features appropriate to the particular scenario they had been given. Other studies have shown that the presence of a theme predicts what specific items, in a set of information items, will later be accurately recalled (Bower, Black, & Turner, 1979; Frederiksen, 1975; Rumelhart, 1975; Thorndyke, 1977).

In sum, then, the reseach reviewed here provides strong evidence that either imposing category structure on stimulus configurations or encountering stimulus configurations that are good matches to existing categories increases overall recall, especially the recall of category relevant information. The following section summarizes the key ideas presented regarding attention, categorization, and recall. The focus of this summary is directed at the ways in which these cognitive variables may operate to affect the decision processes involved in a performance evaluation task.

Summary

The preceeding discussion indicates that the processing of information involves scanning the environment, selecting items to attend to, taking in information about those items, and storing it in some form so that it can be retrived for later consideration. To select the information that is useful and to process it quickly and efficiently, the perceiver needs selection criteria and guidelines for processing. Personal hypotheses about how the world works (based upon individuals' categorization systems) provide such criteria by "telling" the perceiver what data to look for, how to interpret the data that are found, and what information will be stored for later recall.

A crucial question concerning the application of these ideas to person perception becomes: how do perceivers classify stimulus people into categories? The following scenario is offered in order to explain the inherent relevance of cognitive information processing to a performance evaluation task. Consider, for example, a supervisor who is asked to evaluate a sales employee in terms of interpersonal skills exhibited with customers. The supervisor must first recall events from the past which were attended to and thus incorporated into his/her "theory" of the employee in question. The previously reviewed research indicates that it should be easier to recall examples of behaviors to justify a particular interpersonal skill rating if that category (dimension) had been used by the supervisor in observing his/her personnel. However, if no such classificatory basis for identifying behavior had been used by the supervisor to begin with, the recall cue of "interpersonal skills with customers" should provide little, if any,

utility in facilitating recall of employee performance on that particular dimension.

It can thus be seen how attention, categorization, and recall are intimately related. For example, without relevant categorization, attention to relevant cues may be nothing more than random or unconscious. Without meaningful categorization, recall may be impaired. Further, the importance of these processes has direct implications for how people should be trained to observe and evaluate the performance of others. Before a more complete discussion of these implications is undertaken, the next section reviews and critiques previous attempts to train individuals to conduct error-free performance assessments of others.

Rater Error Training: Error versus Accuracy

Rater Training Programs

As previously mentioned, the general assumption underlying most previous rater training programs is that certain rating distributions are ipso facto more desirable than others. For example, ratings at about the same level across dimensions and within ratees are considered an indication of halo error, and raters are encouraged to spread their ratings out for the various dimensions when evaluating others. Similarly, negatively skewed distributions are considered an indication of leniency error, and raters are encouraged to conform more closely to a normal distribution. More specifically, in a very detailed training program (Borman, 1979), ratee performances were shown on videotape and 123 student trainees rated them. Ratings were then placed on a flip

chart and rating distributions were compared and errors discussed. In a much simpler version of this same type of training, Borman (1975) defined halo error and presented a rating distribution showing the error. The training consisted of no more than a lecture advising 90 low- and middle-level managers not to cluster their ratings across dimensions. There was no videotape of performance to use as a criterion in this program. Similar training strategies focusing on the presentation of certain rating distributions as an indicator of error include those developed by Bernardin (1978), Brown (1968), Ivancevich (1979), Levine and Butler (1952), Warmke and Billings (1979), and Bernardin and Boetcher (Note 2). A major problem with this approach to training, however, is that the reseachers did not seem to consider whether or not a skewed distribution, for example, might in reality accurately reflect the performance of certain employees.

With respect to the effectiveness of this type of training in decreasing various errors, the results are inconsistent. Rating errors have successfully been reduced using Borman's (1975) 5-minute lecture to managers, though lectures to student raters failed to produce similar results (Vance, Kuhnert, & Farr, 1978). Longer lectures have produced reduction in halo error (Bernardin, 1978; Brown,1968) but did not improve foreman's administrative ratings (Levine & Butler, 1972). Similarly, discussion groups focusing on rater errors have proven successful for reducing leniency after 90 minute sessions (Levine & Butler, 1952) but have failed to reduce halo in 2-hour versions (Warmke & Billings, 1979). The only training method to produce consistent decreases in rater errors has been a workshop method developed by Latham

et al (1975), which provides participants with an opportunity to practice observing and rating actual videotaped ratees. This technique has been shown to sharply reduce contrast, halo, similar-to-me, and first impression errors.

Somewhat different approaches to training have similarly produced successes and failures. Bernardin and Walter (1977), for example, found that students who kept diaries of their instructor's teaching performance produced ratings with less leniency and halo than students who had not kept diaries, even though both groups received rater error training. In another study, Taylor and Hastman (1956) found that a treatment in which individual attention was given to supervisor raters during the rating task resulted in less halo.

Unfortunately, all of the studies just reviewed are plagued by one or more deficiencies limiting the usefulness of their results (Spool, 1978). First, the focus on rater behavior in terms of error has left the question of accuracy largely unaddressed. This state of affairs is the result of the general assumption that error and accuracy covary negatively and hence, decreasing error should logically increase accuracy. However, this assumption has been questioned by recent research (reviewed below) which indicates that error reduction does not affect accuracy in the anticipated manner. A second limitation of the previous training programs is the lack of attention to an appropriate theoretical basis to serve as guidance for rating training efforts. Some researchers (e.g., Borman, 1978; Kane & Lawler, 1978; King, Hunter, & Schmidt, 1980; Landy & Farr, 1980) have gone so far to posit that performance appraisals and performance appraisal training programs are

unlikely to improve until an adequate theoretical basis for these processes has been developed and tested.

The following section further addresses these limitations. In light of recent research which questions the prevailing error/accuracy covariation assumption, it is first argued that only focusing on error is too limited for increasing the accuracy or validity of raters' observations. Previous training efforts are then assessed in terms of Feldman's (1981) rating process model. Based upon the implications from this cognitive perspective, a rationale is developed concerning why strategies to reduce error are largely insufficient for improving accuracy. This is based on the fact that previous training methods have not directed attention to and facilitated appropriate categorization and recall of relevant employee behaviors, which are central to effective evaluation procedures (Latham & Wexley, 1981; Smith & Kendall, 1963).

Error versus Accuracy

The assumption that error and accuracy covary negatively has been questioned by four recent studies that used varying approximations to true scores as criteria and were thus able to assess rating accuracy. In the first study, Borman (1977) developed normative true scores for job performance dimensions and used Cronbach's (1955) differential accuracy score to operationalize rating accuracy (which was the correlation between normative true scores and subjects' ratings). Scores reflecting halo, leniency/strictness, and restriction in range errors were also computed. The results indicated that although accuracy was not substantially related to any of the errors (*r's* = .12 to .18),

higher halo seemed to moderately covary with higher accuracy. The covariation between accuracy and the other two errors was unclear in that both positive and negative correlations resulted across different jobs. In a second study, which was an extension of the first, Borman (1979) used a variety of rating formats and training procedures to evaluate their effects on halo error and rating accuracy. The results of this research showed that training significantly reduced halo, but did not improve accuracy. Two other studies (Berman & Kenny, 1977; Warmke, 1980) similarly revealed a relatively low relationship between halo and accuracy, and equally unclear results concerning the direction of their covariation.

Although these four studies were not primarily designed to assess error/accuracy relations, taken together they suggest a paradox, at least with respect to halo and accuracy. Cooper (1981) has estimated the halo/accuracy relationship by summarizing the data from the four studies just presented. Halo and accuracy were shown to share a median of 8 percent of the variance, but the direction was opposite to the prevailing negative covariation assumption (i.e., higher halo and higher accuracy modestly covaried). Research investigating the covariation between accuracy and other rating errors is so limited that conclusions must, as yet, remain speculative. However, one conclusion that can be drawn is that additional research investigating the relationship between error and accuracy is clearly warranted.

To summarize, then, the research reviewed here seems to suggest that the basic assumption underlying the development of rater training programs to date (that decreasing error will increase accuracy) is

questionable. Further, the failure to investigate rating accuracy or validity in program evaluation is unfortunate because accuracy is the crucial criterion for judging the quality of performance ratings. More critical, however, is the possibility that error reduction training does not significantly increase accuracy, leaving the utility of previous rater training efforts seriously in doubt.

A plausible explanation as to why strategies to reduce error may not increase accuracy is suggested by Feldman's (1981) rating process model. First, it seems that previous training programs have directed trainees' attention away from the observation of relevant employee behaviors and toward monitoring their own rating behavior in terms of "errors." This seems especially true for those programs which used drawings of rating distributions on flip charts as their focal training tool. Further, error reduction training has not provided raters with an appropriate schema for observing and interpreting behavior, hence, they have done nothing to facilitate accurate recall of raters' observations. The following section further draws upon the rating process model and its implications for the development of an approach to rater accuracy training.

Rater Accuracy Training

Rating Process Implications for Accuracy Training

Feldman's (1981) rating process model provides a useful theoretical basis from which rater accuracy training can be developed. This model states that the interdependent processes of attention, categorization, and recall play a vital role in performance evaluation. Further,

although there are sometimes profound individual differences in the stimuli attended to, the way they are categorized, and what information is recalled, the previously reviewed research on attention, categorization, and recall indicates that these processes can be externally influenced. For example, it has been shown that people can be cued to become consciously aware of certain stimuli in their sensory field, and that this attention increases individuals' ability to accurately recall the information attended to (Averbach & Coriell, 1961; Eriksen & Collins, 1969; Lawrence, 1961; Treisman & Geffen, 1967; Treisman & Riley, 1969). It has also been shown that individuals' category systems direct their attention to particular stimuli and provide the basis for interpreting it (Marcus, 1977; Minsky, 1975; Tesser, 1978). Finally, it has been shown that meaningful category systems can be imposed on people from external sources, and that these facilitate the recall of category relevant information (e.g., Bower, Black, & Turner, 1979; Potts, 1972; Zandy & Gerard, 1974).

According to this view of the rating process and the research supporting it, certain implications for what kinds of training might increase the accuracy of performance ratings are suggested. First, training focused on standardizing the behaviors attended to or consciously looked for would be important. Second, the model implies the importance of teaching raters a common way of defining, organizing, interpreting, and hence recalling the relevant behaviors that are observed (e.g., a common frame-of-reference for categorizing different job behaviors and their effectiveness levels should be provided to raters). The model implies, then, that in order to increase the

accuracy of performance ratings, rather than (or possibly in addition to) focusing on errors, attention should be focused on training in behavior observation and creating or imposing a type of organizing schema to facilitate the storage and recall of relevant observations.

It is interesting to note that these implications are consistent with the contention of Landy and Farr (1980) that raters develop common frames-of-reference for rating job performance. These authors further state that rating quality should be improved if appraisers carefully attend to the performance requirements of the job when rating others. Preliminary support for this notion can be found in the industrial/organizational psychology literature which shows that the use of particular job relevant categories influence the quality of the interview decisions that are made. Specifically, Langdale and Weitz (1973) and Weiner and Schneiderman (1974) found that when available to interviewers, job information was more readily used in their decisions, and that it served to decrease the effects of irrelevant information for both experienced and inexperienced interviewers. Thus, being more familiar with the requirements of the job seemed to help focus the interviewers attention on those applicant qualifications which were more relevant to the person-job fit (Landmark Schmitt, 1976). While this conclusion is supported by limited research involving organizational decision processes, results of research from other literatures focusing on the training of behavior observers have generally supported the promise of this approach (Jecker, Maccoby, & Brietrose, 1965; Wahler & Leske, 1973).

Design of Rating Formats for Accuracy Training

Performance evaluation processes which capitalize upon the major elements of Feldman's model and the suggestions of Landy and Farr (1980) are not entirely unrepresented in the fields of Industrial Psychology and Organizational Behavior. It must be noted, however, that the originators of these few approaches have not consciously acknowledged their theoretical consistency with the cognitive psychology area in general. Nonetheless, Behavioral Observation Scales (BOS; Latham & Wexley, 1977, 1981) and Behaviorally Anchored Rating Scales (BARS; Smith & Kendall, 1963) seem to be constructed and used in a manner which is consistent with the above implications in many ways. First, the specificity of the behavioral examples of these instruments could be used to cue raters' attention to the relevant performance requirements of the job. Second, the dimensionality of these types of instruments seem analogous to the structure of cognitive categories. Specifically, BARS and BOS are characterized by several job performance dimensions, each of which is further defined by examples of specific employee behaviors, and the degree to which these are effective or ineffective. Hence, on a lower level of abstraction, the organization inherent in personal category systems is replicated in these instruments because the general performance dimensions are similar to broad cognitive category domains, and the employee behaviors (which may serve to facilitate the development of dimensional prototypes of effective and ineffective employees) represent more specific information comprising these "categories."

It seems that these instruments would act to impose a common schema or categorization system on raters whereby relevant employee behaviors could be similarly defined, organized, interpreted, and hence, accurately recalled. However, the results of many format comparison studies have not shown that any one type of scale is best. For example, although the BARS format is an elegant strategy for developing performance rating scales, little if any psychometric superiority has been evidenced by this approach over others (Bernardin, 1977; Dunnette & Borman, 1978; Schwab, Heneman, & DeCotiis, 1977). In fact, certain types of scales have outperformed BARS at times (Bernardin, Alvares, & Cranny, 1976; DeCotiis, 1977), although this could be due to variation in scale development and scoring procedures not entirely consistent with the original BARS methodology (Bernardin et al, 1977; Borman, 1979). Comparative studies involving BOS are too limited at this time to warrant any conclusions regarding their superiority (or lack of) over other formats. In sum, however, no clear-cut advantage has been found for any one performance rating format.

A potential reason why such behavioral formats have not generally been shown superior is that merely instructing people to use a certain format (category system) may not be sufficient to really impose that category structure on their thinking. The typical practice in an organization that is developing a new performance appraisal system is to include a small subsample of individuals familiar with a job who then aid in developing the performance appraisal dimensions and behaviors. The participation of these individuals could be expected to facilitate their acceptance and use of the category system they mutually conceive

of as correct. However, the majority of people who would then be asked to use the new format but who did not participate in its development might be less accepting of the new category system. This relative lack of acceptance may be due to simple unfamiliarity with the category system (general dimensions of job performance) and/or the lack of awareness of relevant behaviors that attend upon it.

Previous research has shown that people do tend to use category systems that are familiar to them (Wyer & Scrull, 1980). Thus, although BARS and BOS formats provide raters with the ability to facilitate rating accuracy, persistence in the use of previously learned category systems may represent a lack of awareness and/or a lack of requisite motivation to take advantage of the new formats. Indeed, the majority of raters are most likely unaware (on a conscious level) of the category systems they use to evaluate others. Further, if this awareness does exist but raters are not convinced that their personal, familiar categorization processes are inadequate, there is little reason to expect that they will embrace a newly imposed system.

Recall, however, that previous laboratory research has shown that individuals are willing to attend to stimuli to which experimenters have directed their attention and that recall can be stimulated through the use of an imposed category system. It seems reasonable to expect that individuals may be more willing to accept an imposed category system in a laboratory rather than a field setting. If, however, it can first be shown that subjects can be successfully trained to increase the accuracy of performance appraisals using BARS/BOS and the implications from cognitive psychology, further research in the field which focuses on the

unique implementation problems of that setting can be attempted.

Based upon the arguments thus far presented, it seems reasonable that the use of actual behavioral instruments as a training tool along with focusing rater attention to the particular job performance dimensions and their corresponding behavioral examples should promote the delevopment of appropriate category systems for observing employee behavior, provide raters with examples of what constitutes effective and ineffective behaviors on each performance dimension (category), and thus, facilitate accurate recall of relevant job related evaluation criteria. In sum, this type of training would not only develop categories more in keeping with the actual job requirements, but, the prototypes developed would be based strictly upon relevant employee behaviors. Thus, irrelevant characteristics such as sex, race, attractiveness, etc. would not be included in the category attributes. It seems logical, then, that this strategy would allow behaviors to be noticed, stored, and recalled in a more useful manner.

Objectives of the Present Research

The purpose of the present research was to evaluate the differences in rater training as a function of orientation of the training program. Also of interest was the assessment of any potential differences in program effectiveness for different job performance dimensions. Specifically, Rater Error Training (RET) was compared to Rater Accuracy Training (RAT) which focused on providing raters with an appropriate categorization scheme for attending to and recalling relevant employee behaviors. Further, the present research employed a completely crossed

experimental design, whereby some subjects received both forms of training, others received either error or accuracy training, and some received no training. For five behaviorally defined performance dimensions, these treatments were assessed in terms of their effects on rating errors (halo and leniency) and accuracy in ratings of videotaped ratees.

METHOD

This section describes the the subject group, research design, and procedures for conducting the experiment. Also presented are the two training programs and the development of the videotapes and rating formats.

Subjects

Participants in the study were 108 undergraduate students enrolled in an introductory industrial/organizational psychology course at a large midwestern university. The total sample consisted of 58 females and 50 males. Their mean age was 20.64 years, and approximately half (N = 57) reported having previous experience with performance appraisal (either rating the performance of others or having their performance rated). Students were randomly assigned to one of four experimental groups described under the Experimental Design section below (n=27 per group). Although the use of student raters raised potential concerns with generalizability to a true rater population, it has been shown that employment decisions made by students in laboratory settings are similar to those made by professional interviewers (Bernstein, Hakel, & Harlan, 1975; Schmitt, 1976). Thus, as well as adding credence to the use of college students as raters, this finding also suggests that low generalizability may not be a particularly salient problem in the

present study.

Experimental Design

A 2 x 2 completely crossed factorial design was used in the present research. The first factor, RET, consisted of two conditions: those who received error training and those who did not receive such training. The second factor, RAT, similarly consisted of two conditions: those who did and did not participate in the accuracy training. Those subjects who received both RET and RAT did not, however, participate in the complete version of each training program (described below). This was not possible for both practical and theoretical reasons. First, separate presentations of RET and RAT would have necessitated three hours of training, thereby doubling the duration of RET/RAT group's training time. Second, this procedure would also have provided students in the RET/RAT condition with twice as much practice using the rating scales and becoming familiar with their behavioral examples and definitions. If, then, the results revealed that the RET/RAT appraisals were more accurate and/or contained less error than the other conditions, the question of whether the results were due to the need for both types of training or whether they were merely a function of increased laboratory time and/or practice would have remained.

In order to prevent such problems with subsequent interpretation of the data and to insure equivalence of the training treatments, the RET/RAT program was limited to a one and one-half hour session. This was accomplised by giving students feedback on the accuracy of their ratings as well as by discussing various rating errors and how they

might be alleviated. Hence, students in the RET/RAT condition were trained by incorporating the major elements of each individual program without, however, requiring an increase in total training time or additional practice with the instruments. In summary, the following experimental conditions were compared in the present study: (1) RET and RAT; (2) RET only; (3) RAT only; and (4) No Training (see Figure 2 for a diagram of the research design).

Procedure

Two weeks before the data collection was to begin, the research project was explained to the entire class. Students were told that the study involved performance appraisals and that they would be asked to rate videotaped performances of several managers talking with a problem subordinate. Participation in the research was voluntary. However, extra credit points were given to those individuals who agreed to be involved in the study.

Subjects placed in training treatments attended their respective programs within the next two weeks. In order to keep group sizes manageable, 12-15 students participated in each session. Immediately following the training program(s), subjects observed and rated videotaped managers. Those subjects in the No Training condition were asked only to observe the videotapes and make their ratings following each manager's performance. After the experiment was completed and the results analyzed, the subjects were fully debriefed.

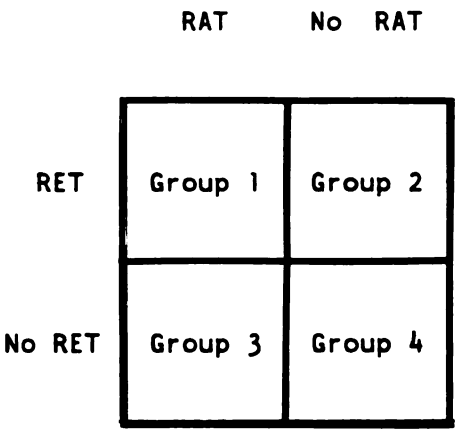


Figure 2. Experimental Design

Videotape and Rating Scale Development

This section presents the procedures as described by Borman (1977) for developing the videotapes and rating scales used in the present research.

Rating Scales

Performance rating scales for a manager talking with a problem subordinate were developed using behavior scaling methodology (Smith & Kendall, 1963; Dunnette, 1966). Seven-point rating scales were used to represent the following seven dimensions of the manager's job:

1. Structure and control of the interview.
2. Reacting to stress.
3. Obtaining information.
4. Resolving conflict.
5. Developing the subordinate.
6. Establishing and maintaining rapport.
7. Motivating the subordinate

Each dimension was defined by both an overall defining statement as well as by scaled behavioral anchors describing the seven different effectiveness levels (see Appendix A for these scales).

Generating Intended "True Scores" for Performers

To make the performances as realistic as possible, "intended true scores" with a preset covariance structure were generated. First, two

realistic covariance matrices were formed by asking experts to estimate the "true" means and standard deviations of performance on each dimension and the "true" intercorrelations among dimensions. Profiles reflecting the "correct" covariance structure were then generated for eight ratee performances.

More specifically, five expert judges knowledgeable about the job and the concept of correlation were asked to independently estimate the level of correlation expected between each pair of dimensions when the job is actually being performed. To accomplish this, they used a 1 to 7 scale, where 7 indicated $\underline{r} = 1.00$; 6, $\underline{r} = .67$; 5, $\underline{r} = .33$; 4, $\underline{r} = .00$; 3, $\underline{r} = -.33$; 2, $\underline{r} = -.67$; and 1, $\underline{r} = -1.00$. A descriptive estimate of reliability associated with these judgments was obtained by using an ANOVA procedure to compare the variability in different judges' ratings of the same dimension pairs with total variance in the judgments. The resulting intraclass correlation for these judgments was .81 ($\underline{p} < .01$), suggesting acceptable reliability for the judgment task. Mean ratings (on the 1 to 7 scale) were computed for each dimension pair, and these means were transformed directly to correlation coefficients (e.g., 4.5 was transformed to +.17).

Following a procedure outlined by Naylor and Wherry (1965), the resulting correlations along with dimension means of 4.0 and standard deviations of 1.5 were then used to generate an intended true score matrix for ratees. As an example, presented below are intended performance profiles for two managers:

Performance Dimension	Profile 1	Profile 2
Structure and Control of the Interview.	6.0	2.0
Reacting to Stress.	5.0	3.5
Obtaining Information.	6.0	2.5
Resolving Conflict.	6.0	5.0
Developing the Subordinate.	3.5	3.5
Establishing and Maintaining Rapport.	4.5	6.0
Motivating the Subordinate.	5.0	2.5

The procedures outlined above thus enabled the development of realistic mulitidimensional performance profiles for eight individuals on the managers' job.

Developing and Videotaping Performance

Eight scripts were written depicting 5- to 9-minute performances of a manager talking with a problem subordinate. The scripts reflected the performance levels defined by the intended true scores as closely as possible. Eight different actors played the various manager roles while the same actor played the problem subordinate in all eight performances. Each actor was given explicit instruction and ample preparation time to insure close conformance to the scripts during the videotaping.

Obtaining Final True Scores

Fourteen expert raters were selected to evaluate the effectiveness of each performer. Seven of the raters were graduate students in psychology, and the other seven were practicing industrial psychologists working either for a psychological consulting firm or in the personnel research department of a large manufacturing company. All of the raters were very familiar with the performance demands of the job. The scripts were revised as necessary to reflect the verbal behavior actually depicted in the performances, and raters were asked to study these scripts and the rating scales before coming to the rating sessions.

Experts' ratings were analyzed using an indirect validation approach. Interrater agreement among the 14 experts was computed for each dimension using intraclass correlations. The resulting eight intraclasses ranged from .91 to .98 with a median of .97. Further, correlations between mean expert ratings and intended true scores were all above .70, with a median r = .93. These results indicated considerable agreement between the expert judges and intended true scores. The high interrater agreement obtained for each dimension suggested that the few times that the mean expert ratings did differ somewhat from the intended true scores, the discrepancies were most likely due to the scripts reflecting unintended levels of performance and/or the actors failing to project the intended effectiveness levels. The mean expert ratings (see Table 1) were therefore adopted as the "true scores" for subsequent uses of the tapes.

Table 1. True Scores of Performance

Dimension/Manager	1	2	3	4	5	6	7	8
Structure and Control of the Interview	2.79	2.77	6.92	2.07	3.31	4.54	4.38	3.08
Establishing and Maintaining Rapport	1.50	5.93	3.26	5.00	3.69	5.23	3.08	1.38
Reacting to Stress	3.57	5.00	5.38	4.29	4.46	4.92	5.15	1.85
Obtaining Information	2.36	4.21	6.15	3.43	1.77	5.69	2.69	1.54
Resolving Conflict	2.07	4.07	5.62	5.00	5.69	4.31	2.85	2.08
Developing the Subordinate	2.71	3.07	3.38	2.93	6.08	6.62	4.54	1.38
Motivating the Subordinate	2.29	4.86	4.62	3.71	5.77	6.15	2.77	2.08

Manipulations

Rater Error Training (RET)

Latham et al (1975) developed a training procedure to help managers become aware of problems in rating employee performance and to reduce various rating errors. The major elements of the Latham et al workshop training procedure were used to train student raters to provide more error free performance assessments. The core characteristics of the method include the following:

1. A videotape of a job being performed is first shown to participants.
2. Trainees then evaluate the designated ratee on the videotape using rating scales as provided.
3. Ratings made by participants are placed on a flipchart.
4. Differences between the ratings and reasons for the differences are discussed by trainees.
5. The trainer discusses rating errors made by ratees and how they can be avoided.
6. The group then discusses ways of avoiding or overcoming the error being studied.

This general strategy was followed for the present rater error training. Specifically, subjects were shown two of Borman's eight videotapes during the training, and they were asked to evaluate each manager's performance using the rating scales that appear in Appendix A.

Because two of the tapes were used as part of the training program, criterion ratings were obtained only on the remaining six videotapes. Subsequent to rating each manager, the trainer discussed subjects' ratings in terms of rating errors such as halo, leniency, central tendency, and contrast effect (see Appendix B for a detailed description of RET). The emphasis of this training was thus focused on producing error-free performance ratings.

Rater Accuracy Training (RAT)

Based on the implications of Feldman's (1981) rating process model as discussed in the introduction, RAT focused on facilitating the development of a common categorization system based upon important job dimensions (which are further defined by specific behaviors) for observing ratee performance. Specifically, those who received the RAT program were first lectured on the multidimensionality of most types of jobs and the need to pay close attention to employee performance in terms of these dimensions. Participants were then given the actual scales they would be using to rate the managers. After discussing the general definitions of each dimension and the behavioral anchors that corresponded to different effectiveness levels, subjects practiced using the rating scales by rating the same two videotaped managers that were used in RET. After the group rated each of the tapes, they discussed their ratings and received feedback on their accuracy. This exercise served to increase the group's attention to the performance dimensions they used to evaluate the managers, and it also served to illustrate various effectiveness levels within each category (see Appendix C for a

detailed description of RAT).

In sum, by using 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, the development of appropriate category systems for observing ratee behavior was promoted. Further, this strategy was expected to provide raters with behavioral examples of what constituted effective and ineffective behavior on each performance dimension (category). This development of categories based on actual job requirements was, in turn, hypothesized to facilitate more accurate recall and evaluation of relevant performance criteria.

Summary of RET and RAT Programs

Based upon the previous discussion of RET and RAT, it can be seen that both training programs were designed to elicit active trainee participation and to provide raters with practice and feedback on their judgments. Also, both training programs used the same two videotapes and the actual rating scales to train participants. Further, in order to control for variance due to differences in the amount of actual training time, the programs were each developed to last approximately one hour and one-half hours. In summary, RET and RAT were identical with respect to their training components (i.e., practice and feedback), training tools, and duration. Hence, any differences between the experimental groups could more confidently be attributed to the focus of the training itself (error or accuracy), rather than to differences in variables extraneous to the present research question (e. g., training

time, components of training, etc.).

Pretesting of Training Programs

Prior to the experimental treatments, the training programs were each pretested with two groups of 10 to 15 students. These pretests were performed to provide the trainer with practice conducting the sessions and also to discover any potential problems with the programs so that modifications could be made prior to the actual research. While the original conceptualization of RET required no major modifications, the RAT pretests revealed the need for various changes. Based upon interviews with pretest subjects as well as the results of preliminary data analyses, it was evident that subjects did not have enough training time to assimilate the amount of information associated with seven performance dimensions. It was impossible to extend the training sessions because of practical limitations regarding the amount of experimental time available from subjects. Hence, two of the seven dimensions, i.e., Reacting to Stress and Obtaining Information, were deleted from the rating scales in all experimental conditions. These particular categories were excluded because subjects reported difficulty in differentiating the effectiveness levels within them. Given that clearly defined behavioral dimensions were a prerequisite to the accuracy training proposed here, the inclusion of obviously ambiguous dimensions would not have facilitated a reasonable comparison of the techniques.

Dependent Variables

Accuracy

One measure of accuracy was calculated using an approach similar to that used by Bernardin & Pence (1980) and Rush, Phillips, & Lord (1981). This measure, Distance, assessed how close the subject was to the mean true score for each of the five dimensions across ratees. The formula for calculating Distance (DIST) is presented in the equation below.

$$A = \sum_{r=1}^{R} (D/R)$$

where:

A = accuracy across ratees for each dimension.

R = number of ratees (6).

D = absolute difference of the observed score from true score.

For each subject, this analysis resulted in five mean deviation scores across ratees, with lower deviations indicating higher accuracy.

Accuracy was also assessed using Cronbach's (1955) differential accuracy (DA) measure. The DA provided accuracy scores for each rater on each performance dimension by correlating the rater's ratings of the six videotaped target persons on a dimension with mean true scores provided by the expert judges. The Fisher r-to-z transformation was then applied to each DA correlation. Thus, for each subject, these analyses resulted in five z scores across ratees within dimensions (DA), with higher scores indicating higher accuracy.

Halo

Halo is conceptualized as the tendency for raters to restrict their ratings of a target person across job dimensions. Operationally, halo has been discussed in terms of standard deviations across dimensions within ratees (e.g., Borman, 1977). In addition to this measure of halo, analyses were also performed for a second halo measure.

First, to test for differences in halo defined in terms of standard deviations, a standard deviation (SD) was computed for each target ratee, thereby reflecting the spread in those ratings across dimensions. A low standard deviation across dimensions indicated high halo and higher standard deviations indicated lower halo. Because of the nonnormal distribution of standard deviations, a logarithmic transformation of the variances was performed before averaging these scores. O'Brien (1978) has recently shown that tests for determining differences in variances using the logarithmic transformation were both robust and powerful.

The second measure of halo (HALOCORR) was calculated in the following way: A correlation matrix was computed between the five dimensions for each subject's ratings of the six ratees. These dimension intercorrelations were then subtracted from the true dimension intercorrelations, yielding 10 difference scores for each subject. Before subtracting the matrices, all correlations were transformed to z scores using Fisher's r-to-z transformation. The difference scores for each subject were then averaged, providing a mean measure of the difference between the true and observed intercorrelations across

dimensions. To the degree that this average deviated from zero in a positive direction, the subject's ratings were less correlated than the true ratings. To the degree that this average deviated from zero in a negative direction, greater halo was evidenced.

Leniency

Leniency (LEN) was assessed for each subject by computing the mean ratings for each dimension across the six ratees. This resulted in five leniency scores for each rater. The mean true scores for each dimension were then subtracted from the observed mean ratings, with greater distance (i.e., larger positive difference scores) indicating greater leniency.

Data Analyses Procedures

For each of the two accuracy measures (i.e., DIST and DA), the experimental groups were compared with a 2 x 2 x 5 (RET x RAT x DIM) fixed-factor analysis of variance (ANOVA) with repeated measures on the dimension factor. This design enabled not only the evaluation of treatment main effects and interactions, but it also allowed the assessment of dimension effects as well as dimension x training interactions. For each of the Halo measures (i.e., SD and HALOCORR), a 2 x 2 ANOVA with RET and RAT as fixed factors was performed to assess differences among the experimental groups. Finally, a 2 x 2 x 5 ANOVA with repeated measures on the last (i.e., dimension) factor was used to assess training and dimension effects and interactions for the leniency (LEN) measure. Table 2 presents a summary of the five dependent

variables, how they were calculated, and the design used to analyze each.

Table 2. Summary of the Dependent Variables

Variable	Definition	Design
Accuracy		
DIST	Average distance from true scores for each of the five performance dimensions	2 x 2 x 5 ANOVA (RET x RAT x DIM) with repeated measures on DIM
DA	Correlation between the true and observed ratings for each of the five dimensions	2 x 2 x 5 ANOVA (RET x RAT x DIM) with repeated measures on DIM
Halo		
SD	Average standard deviation within ratees	2 x 2 ANOVA (RET x RAT)
HALOCORR	Average distance between true and observed dimension intercorrelations	2 x 2 ANOVA (RET x RAT)
Leniency		
LEN	Difference between true and observed means for each of the five dimensions	2 x 2 x 5 ANOVA (RET x RAT x DIM) with repeated measures on DIM

RESULTS

Relationships Between Accuracy and Rating Errors

The means, standard deviations, and intercorrelations for subjects' accuracy scores and scores for the two rating errors are presented in Table 3. Also shown are the correlations between the dependent variables, sex, age, and previous experience with performance appraisals.

The relationship between the two measures of accuracy (i. e., distance from true scores - DIST and differential accuracy - DA) was substantial ($r = -.79$, $p < .05$). The negative correlation indicated that smaller absolute distances from the true scores were associated with higher correlations between dimension true scores and observed scores. There was also a relatively large amount of overlap between the two halo measures ($r = -.77$, $p < .05$). Specifically, those individuals who had larger deviations within ratees (i. e., less halo) had dimension intercorrelations that were lower than the true dimension intercorrelations, whereas those with smaller SD measures had dimension intercorrelations that were greater than the true dimension intercorrelations (i. e., higher halo). Although the correlation between the two accuracy measures and between the two halo measures was high, separate analyses were conducted on each measure so that the present analyses would be comparable with previous research.

Table 3. Means, Standard Deviations, and Intercorrelations of Variables[a]

Variable[b]	Mean	SD	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. DIST	1.13	.62								
2. DA	1.01	.25	-.79							
3. SD	.05	.24	.07	.04						
4. HALOCORR	.27	.36	.24	-.15	-.77					
5. LEN	.15	.47	.27	-.07	-.15	.11				
6. SEX	1.54	.50	-.05	-.05	-.11	.04	.14			
7. AGE	20.64	2.06	.02	-.05	.09	-.13	-.01	-.13		
8. EXPER	1.53	.50	.07	.08	.09	-.04	.14	.05	.00	

[a] $\underline{r} > .15$, $p < .05$

[b] DIST = accuracy measured as distance from true scores; DA = accuracy measured by differential accuracy; SD = halo as the standard deviation within ratees; HALOCORR = halo measured as the average difference between true and observed dimension intercorrelations; LEN = leniency measured as the average difference between observed and true means.

There was virtually no relationship between halo measured in terms of SDs and either the DIST or the DA accuracy measures. Low but statistically significant correlations resulted between the two accuracy measures and HALOCORR. Both of these correlations were negative, indicating that lower accuracy was associated with positive deviations from true dimension intercorrelations (i. e., higher halo). While this finding may at first seem to support the notion that error and accuracy covary negatively, it must be remembered that the HALOCORR measure was based upon the true dimension intercorrelations. Hence, this particular measure of halo was not consistent with previous operationalizations of the error that did not involve the true scores (e. g., Bernardin & Pence, 1980; Borman, 1975, 1979).

Leniency (LEN) was not related to the DA measure of accuracy but was significantly related to the DIST measure (r = .27, p < .05). Specifically, more accurate ratings (smaller distances from true scores were associated with negative deviations from the true means, while leniency (positive deviations from the true means) increased with inaccuracy. Again, however, it must be noted that the leniency measures used here were based on the true means. There was no relationship between leniency and either of the halo measures.

Training Effects on Accuracy

Distance from True Scores

A 2 x 2 x 5 ANOVA (RET x RAT x DIM) with repeated measures on the last (i. e., dimension) factor was performed to assess the effects of

training on the DIST measure of accuracy. This design also enabled the assessment of dimension effects as well as training x dimension interactions. The results of the ANOVA (shown in Table 4) revealed a significant main effect for RAT. Inspection of the means in Table 5 indicated that individuals who participated in RAT had significantly more accurate ratings that those who did not receive RAT. More noteworthy, however, was the significant RET x RAT interaction. Analysis of the mean data (presented in Figure 3) suggested that RAT alone produced the most accurate ratings while the no training group was least accurate. Tukey tests specifically revealed that RAT alone or RET/RAT together yielded ratings with higher accuracy than no training or RET alone. Further, there were no differences in accuracy between the no training and RET alone conditions.

A main effect for DIM and two significant training x dimension interactions were observed. The significant RAT x DIM interaction (see Figure 4) revealed differences in the effectiveness of RAT on the appraisal dimensions. With each dimension fixed, evaluations of the simple main effects for designs with repeated measures (Winer, 1971) showed RAT to significantly increase accuracy on only three (i. e., Structuring and Controlling the Interview, Resolving Conflict, and Developing the Subordinate) of the five dimensions. Further analyses for only the RAT group revealed that Structuring and Controlling the Interview ($\bar{x}$ = .72) was rated more accurately than all other dimensions, while Establishing and Maintaining Rapport ($\bar{x}$ = 1.24) was rated the least accurately. The same analysis conducted for the NO RAT group showed that Resolving Conflict ($\bar{x}$ = 1.45) was rated less accurately than

Table 4. Results of the Analysis of Variance for DIST

Effect	df	F	ω^2
RET (A)	1	.12	
RAT (B)	1	52.25*	.30
A x B	1	10.69*	.06
Subjects x A x B	104	(.20)	
DIM (C)	4	12.76*	.06
A x C	4	4.11*	.01
B x C	4	7.88*	.03
A x B x C	4	1.82	
Subjects x A x B x C	416	(.12)	

Note. Numbers in parentheses are the mean square error associated with the F tests directly above them in the table.

* $p < .05$

Table 5. Means and Standard Deviations of DIST

Variable	NO RET	RET	Totals
NO RAT	1.33 (.25)	1.22 (.22)	1.27 (.24)
RAT	.92 (.16)	1.06 (.15)	.99 (.17)
Totals	1.13 (.29)	1.14 (.21)	1.13 (.25)

Note. Numbers in parentheses = SDs.

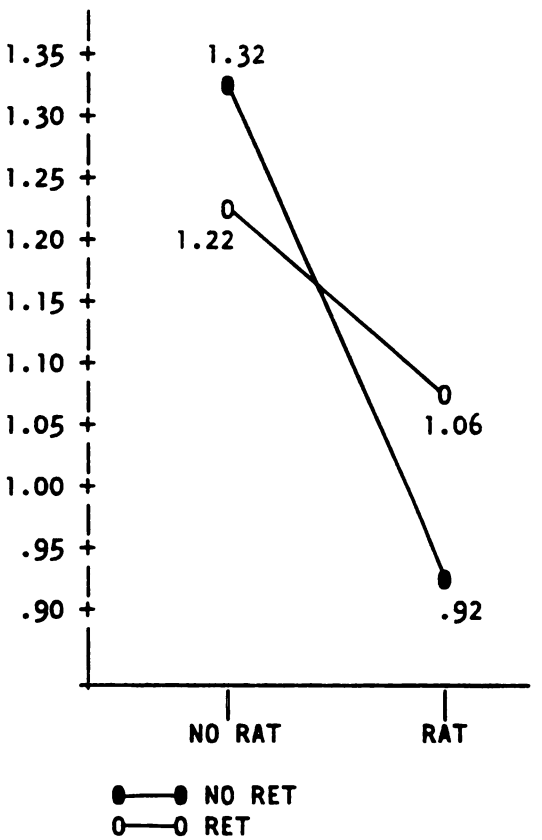
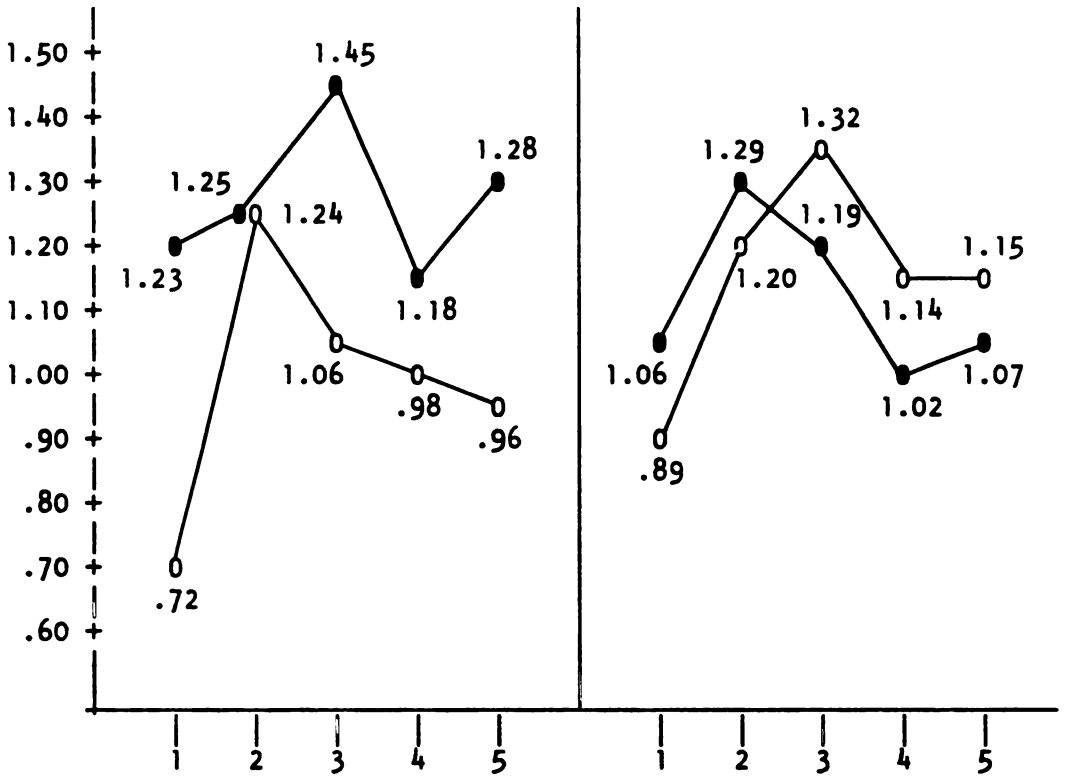


Figure 3. Mean Data (DIST) RET x RAT Interaction



1 = Structuring and Controlling the Interview; 2 = Establishing and Maintaining Rapport; 3 = Resolving Conflict; 4 = Motivating the Suborainate; 5 = Developing the Subordinate

●—● NO RAT
O—O RAT
●—● NO RET
O—O RET

Figure 4. Mean Data (DIST) for DIM x Training Interactions

the remaining four dimensions.

A significant RET x DIM interaction also resulted and is shown in Figure 4. Although an analysis of the simple main effects revealed no significant differences between the dimension means for the two RET conditions, the profiles of these means were different for the RET versus NO RET group. However, further analyses of the means within the two treatment groups did reveal differences in accuracy for particular dimensions. Specifically, for the RET group, Structuring and Controlling the Interview ($\bar{x}$ = .89) was rated more accurately than the other four dimensions, and with the exception of Establishing and Maintaining Rapport ($\bar{x}$ = 1.20), Resolving Conflict ($\bar{x}$ = 1.32) was rated less accurately than the others. For the NO RET group, Establishing and Maintaining Rapport ($\bar{x}$ = 1.29) was rated with less accuracy than all other dimensions except for Resolving Conflict ($\bar{x}$ = 1.19).

Differential Accuracy

A 2 x 2 x 5 ANOVA with repeated measures on the last factor was also performed to assess the effects of training and dimensions on the transformed (r-to-z) DA correlations. Cell means and standard deviations are presented in Table 6, and Table 7 contains the results of the ANOVA as well as omega square values for the significant effects.

A significant main effect resulted for RAT, whereby those who received training had significantly higher correlations between true and observed dimension scores than those who did not receive accuracy training. A significant RET x RAT interaction also resulted for the DA measure (see Figure 5). The nature of this interaction was somewhat

Table 6. Means and Standard Deviations of DA[a]

Variable	NO RET	RET	Totals
NO RAT	.82 (.25)	1.01 (.19)	.91 (.24)
RAT	1.22 (.19)	1.01 (.19)	1.12 (.22)
Totals	1.02 (.30)	1.01 (.19)	1.01 (.25)

Note. Numbers in parentheses = SDs.

[a]Values in the table are based on transformed r-to-z correlations.

Table 7. Results of the Analysis of Variance for DA

Effect	df	F	ω^2
RET (A)	1	.06	
RAT (B)	1	25.69*	.16
A x B	1	27.05*	.17
Subjects x A x B	104	(.21)	
DIM (C)	4	21.12*	.11
A x C	4	3.17*	.02
B x C	4	12.49*	.07
A x B x C	4	.67	
Subjects x A x B x C	416	(.21)	

Note. Numbers in parentheses are the mean square error associated with the F tests directly above them in the table.

* $p < .05$

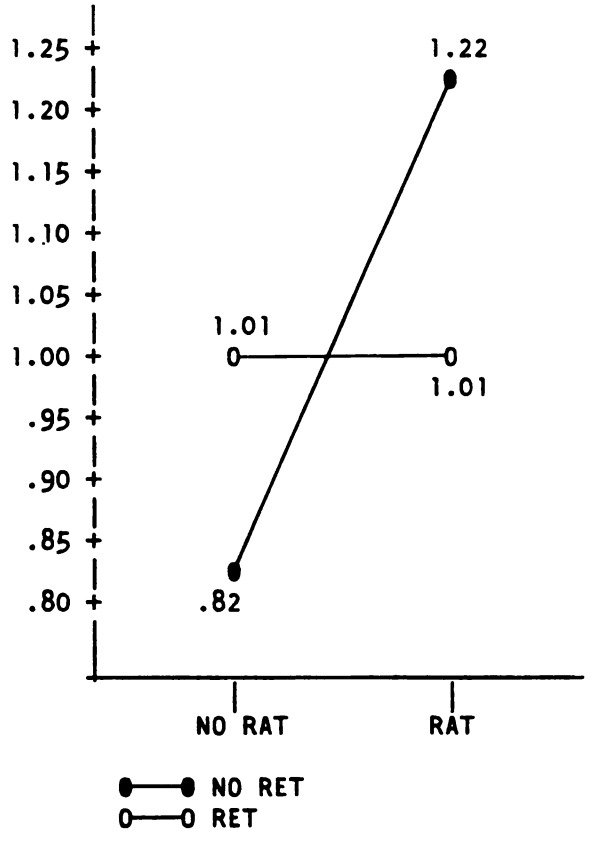
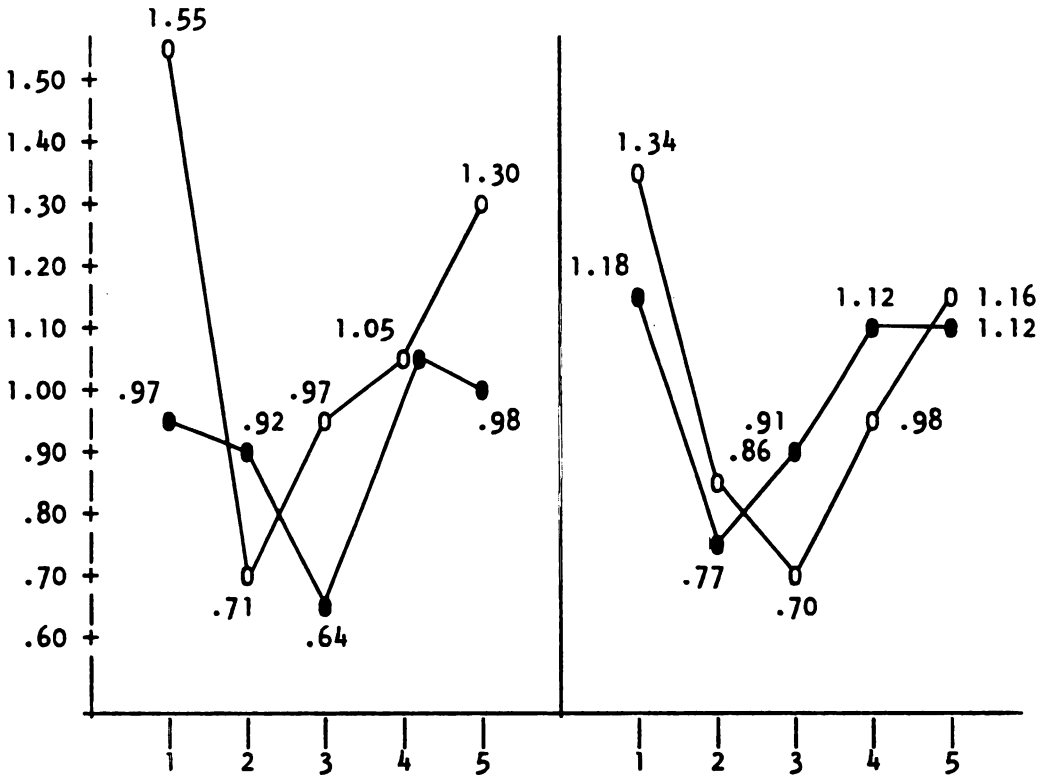


Figure 5. Mean Data (DA) for RET x RAT Interaction

different than the relationship between RET and RAT that resulted with the DIST measure of accuracy. Specifically, Tukey tests revealed that RAT alone was better than any other condition in improving the correlations between observed and true dimension scores. There was no difference in DA correlations from the RET alone versus the RET/RAT condition. However, both RET alone and RET/RAT together produced more accuracy than the no training condition. Interestingly, accuracy was significantly decreased when RET was combined with RAT as compared to the RAT alone condition.

The results of the within subject dimension analysis were essentially the same as those found with DIST. A significant main effect for DIM and significant RAT x DIM and RET x DIM interactions resulted (see Figure 6). Analysis of simple main effects indicated that the nature of these interactions were similar to those found with DIST. Specifically, RAT increased accuracy on only three of the five rating dimensions. Within the RAT treatment, Structuring and Controlling the Interview ($\bar{x}$ = 1.55) was rated more accurately than all other dimensions. The accuracy associated with Developing the Subordinate ($\bar{x}$ = 1.30) was also relatively high, as the ratings on this dimension were closer to true scores than on the remaining three. Finally, Motivating the Subordinate ($\bar{x}$ = 1.05) was rated more accurately than Establishing and Maintaining Rapport ($\bar{x}$ = .71) but not more accurately than Resolving Conflict ($\bar{x}$ = .64). Within the NO RAT group, a significant difference resulted for only one dimension. Specifically, Resolving Conflict ($\bar{x}$ = .64) was rated with less accuracy than the other four dimensions.



1 = Structuring and Controlling the Interview; 2 = Establishing and Maintaining Rapport; 3 = Resolving Conflict; 4 = Motivating the Subordinate; 5 = Developing the Subordinate

●—● NO RAT
O—O RAT

●—● NO RET
O—O RET

Figure 6. Mean Data (DA) for DIM x Training Interactions

With respect to the RET x DIM interaction, although the simple main effects analyses were again nonsignificant, different profiles of accuracy scores across dimensions occurred in the RET versus the NO RET treatment. Within the RET group itself, however, Structuring and Controlling the Interview ($\bar{x}$ = 1.34) was the most accurately rated dimension. Developing the Subordinate ($\bar{x}$ = 1.16) was rated with more accuracy than Resolving Conflict ($\bar{x}$ = .70) and Establishing and Maintaining Rapport ($\bar{x}$ = .86). Finally, Motivating the Subordinate ($\bar{x}$ = .98) was also rated with more accuracy than Resolving Conflict. Within the NO RET group, Establishing and Maintaining Rapport ($\bar{x}$ = .77) and Resolving Conflict ($\bar{x}$ = .91) were rated less accurately than the other three dimensions.

Training Effects on Halo

A 2 x 2 ANOVA was performed to assess training effects on each of the halo measures. Results of the first ANOVA (see Table 8) for the SD measure of halo (i. e., average standard deviation within ratees) revealed significant main effects for both RET and RAT. Further inspection of the mean data presented in Table 9 suggested that error training significantly increased the spread in ratings (i. e., decreased halo). Conversely, accuracy training significantly increased halo (i. e., decreased the standard deviations within ratees). Although both main effects were significant, the omega square value associated with the RET effect was substantially larger than that associated with the RAT effect.

Table 8. Results of the Analysis of Variance for Halo

Effect (SD)	df	F	ω^2
RET (A)	1	53.02*	.31
RAT (B)	1	7.06*	.04
A x B	1	1.09	
Subjects x A x B	104	(.04)	

Effect (HALOCORR)	df	F	ω^2
RET (A)	1	39.49*	.27
RAT (B)	1	.08	
A x B	1	1.87	
Subjects x A x B	104	(.04)	

Note. Numbers in parentheses are the mean square error associated with the F tests directly above them in the table.

* $p < .05$

Table 9. Means and Standard Deviations of Halo

Variable (SD)	NO RET	RET	Totals
NO RAT	.95 (.12)	1.30 (.24)	1.11 (.24)
RAT	.90 (.20)	1.18 (.19)	1.01 (.23)
Totals	.92 (.17)	1.21 (.23)	1.05 (.24)

Variable (HALOCORR)	NO RET	RET	Totals
NO RAT	.50 (.25)	.05 (.37)	.27 (.38)
RAT	.40 (.28)	.11 (.31)	.26 (.33)
Totals	.45 (.27)	.08 (.34)	.27 (.36)

Note. Numbers in parentheses = SDs.

The second ANOVA was conducted on the average difference between observed and true dimension intercorrelations (HALOCORR). The results of this ANOVA (also presented in Table 8) showed only a significant main effect for RET. The means and standard deviations associated with this analysis are shown in Table 9. The dimension intercorrelations of those individuals who participated in error training were closer to the true dimension intercorrelations than for those who did not receive RET. The dimension intercorrelations for the NO RET groups were substantially higher (i. e., more halo) than the true dimension intercorrelations.

Training Effects on Leniency

The main analysis aimed at evaluating training effects on leniency employed the average difference between observed and true means within dimensions in a 2 x 2 x 5 ANOVA, with RET, RAT, and DIM (repeated measures) as fixed factors. Results of that ANOVA (see Table 10) indicated a significant main effect for RAT. Evaluation of the means in Table 11 showed that accuracy training yielded mean dimension ratings that were closer to the true means. Those who did not receive accuracy training tended to rate the managers with more leniency, as evidenced by the positive mean deviation score for the NO RAT group.

A significant main effect for DIM and a significant RAT x DIM interaction also resulted (see Figure 7). Tests of simple main effects revealed that RAT was effective in reducing leniency only with respect to Structuring and Controlling the Interview and Establishing and Maintaining Rapport. Within the RAT group alone, Developing the

Table 10. Results of the Analysis of Variance for LEN

Effect	df	F	ω^2
RET (A)	1	.28	
RAT (B)	1	16.50*	.13
A x B	1	.58	
Subjects x A x B	104	(.96)	
DIM (C)	4	19.70*	.09
A x C	4	1.42	
B x C	4	3.85*	.01
A x B x C	4	.58	
Subjects x A x B x C	416	(.24)	

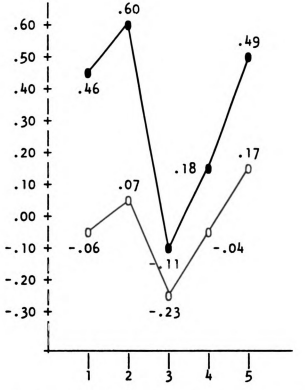
Note. Numbers in parentheses are the mean square error associated with the F tests directly above them in the table.

* $p < .05$

Table 11.--Means and Standard Deviations of LEN

Variable	NO RET	RET	Totals
NO RAT	.27 (.45)	.38 (.45)	.33 (.45)
RAT	-.01 (.40)	-.03 (.44)	-.02 (.42)
Totals	.13 (.45)	.18 (.45)	.15 (.47)

Note. Numbers in parentheses = SDs.



1 = Structuring and Controlling the Interview; 2 = Establishing and Maintaining Rapport; 3 = Resolving Conflict; 4 = Motivating the Subordinate; 5 = Developing the Subordinate

●——● NO RAT
O——O RAT

Figure 7. Mean Data (LEN) for DIM x Training Interactions

Subordinate ($\bar{x}$ = .17) was rated more leniently than Resolving Conflict ($\bar{x}$ = -.23) and Structuring and Controlling the Interview ($\bar{x}$ = -.06). Establishing and Maintaining Rapport ($\bar{x}$ = .07) was also rated with more leniency than was Resolving Conflict. With respect to the NO RAT group, the least lenient (most severe) ratings were associated with Resolving Conflict ($\bar{x}$ = -.11). Less leniency was also observed on Motivating the Subordinate ($\bar{x}$ = .18) as compared to Structuring and Controlling the Interview ($\bar{x}$ = .46), Establishing and Maintaining Rapport ($\bar{x}$ = .60), and Developing the Subordinate ($\bar{x}$ = .49).

DISCUSSION

The results of the present study suggest that rating accuracy can be improved by training individuals in a manner that is consistent with and facilitates human information processing capabilities. Specifically, it appears that the use of an actual behavioral instrument as a training tool had the effect of providing raters with a common frame-of-reference for evaluating ratee behavior. This was not particularly surprising in that the categories one uses are a function of education and experience (Ilgen & Feldman, 1983). Further, by focusing rater attention to the particular effective, average, and ineffective behaviors that corresponded to each rating dimension, trainees were given easily detectable cues of good and poor performance which were hypothesized to enhance the development of their newly imposed, more specialized category systems. Hence, the increases in accuracy found in the RAT group seem to support the notions of several recent researchers who have suggested that the development of job-relevant category systems, along with their implications for the treatment and evaluation of employees (Swann & Snyder, 1980), are the source of valid variance in performance appraisals (Ilgen & Feldman, 1983). Further, the effects of RAT in improving accuracy were evidenced regardless of whether accuracy was conceptualized in terms of distance from true scores (DIST) or the correlations between true and observed

scores on each dimension (DA). This was not unexpected, however, given the substantial degree of overlap between the two accuracy measures.

This study also lends support to previous research (Bernardin & Pense, 1980; Latham et al., 1975) which has shown that individuals can be trained to reduce psychometric errors in their ratings. Specifically, error training reduced halo measured in terms of standard deviations across ratees (SD) and in terms of differences between true and observed dimension intercorrelations (HALOCORR). Error training did not, however, have any effect on leniency. This result may be an indication that the error training used here was simply not as effective as previous training efforts in reducing leniency. For instance, although the Latham et al (1975) workshop procedure was followed, their actual training tapes were not used. However, a main effect for RAT was observed on the leniency measure. Specifically, the ratings of those who received accuracy training were closer to the true dimension means than were the ratings of those who did not receive RAT.

Perhaps some of the most interesting findings, however, concerned the RET x RAT interactions associated with the accuracy analyses. First, when RET was combined with RAT, rating accuracy was significantly decreased as measured by DA. Further, although the mean differences were nonsignificant, the average distance from the true scores was somewhat lower in the RAT alone condition compared to the combined condition. A potential explanation for this result may reflect a potential problem with the RET/RAT training program itself. Specifically, subjects who received both forms of training were presented with twice as much information in the same amount of time as

those who received only RET or only RAT. Further, recall that it was necessary to delete two of the original rating dimensions because one and one-half hours of training was not a sufficient time period to process all seven scales. It thus seems plausible that the RET/RAT subjects may not have been able to efficiently assimilate the amount of information that was required by their particular training program.

However, another plausible explanation for the finding that RET and RAT together tended to decrease accuracy is that when RET was combined with RAT, subjects' attention may have been partially diverted away from the observation and evaluation of relevant ratee behaviors to monitoring their own rating behavior. Concern with avoiding the rating errors discussed during the training session may have to some degree compromised the accuracy of their evaluations. In fact, anecdotal evidence obtained from subjects who participated in the RET/RAT treatment suggests that this may have been the case. Several students reported purposely spreading out their ratings in order to "avoid the errors" when they would have preferred rating particular target ratees more uniformly. Given the present research design, however, it is not possible to ascertain which, if either, of these explanations is valid. Future research aimed at clearly delineating the particular effects of combining the types of training employed here certainly seems warranted.

In terms of comparing the effects on accuracy of error training versus no training, the results are not entirely conclusive. Concerning the DA measure of accuracy, for example, RET significantly increased accuracy as compared to the no training condition. This result is inconsistent with previous research (e.g., Bernardin & Pence, 1980;

Borman, 1979) that has found error training to have no effect on increasing accuracy. On the other hand, and consistent with previous research was the finding of no difference in accuracy (as measured by distance from true scores) between the RET alone and the no training conditions. The question of whether error training is better than no training might be largely dependent upon how one conceptualizes accuracy as well as a function of variations across studies in the particular training strategies and rating scales used. If, for example, our goal is to have ratings that covary accurately with "true scores," then the present results indicate that error training may be better than no training for increasing accuracy. If, however, our goal is to obtain ratings that accurately reflect a ratee's level of performance vis a vis a behavioral rating instrument, then the present results indicate that error training may be ineffective.

Another result to emerge from this study was that the accuracy training employed here was effective on only three of the five rating dimensions. These were: Structuring and Controlling the Interview, Resolving Conflict, and Developing the Subordinate. While only post hoc explanations of this result are possible, it appears that these dimensions may have been more explicitly defined in terms of the particular effective and ineffective behavioral cues corresponding to various performance levels. This explanation seems plausible, especially upon further evaluation and comparison of the behavioral descriptions associated with those dimensions that were affected by RAT versus those that were not. For example, the behavioral cues constituting a "7" on Developing the Subordinate (e.g., "setting up a

specific developmental program for the subordinate," "making worthwhile developmental suggestions such as enrolling in an interpersonal skills seminar or taking the Dale Carnegie course," and "setting up specific days and times to meet and discuss developmental issues and progress") seem less ambiguous than the cues associated with a "7" on Establishing and Maintaining Rapport (e.g., "effectively bringing-out the subordinate's problems through probing but nonthreatening questions" and "discussing the subordinate's problems in a warm and supportive manner"). Similar examples of relatively ambiguous anchors are more prevalent on the two dimensions for which RAT had no effect.

Also of interest was the finding that there were differences observed in accuracy and leniency across the dimensions within particular treatments. This was especially noteworthy because without exception, previous rater training efforts have focused on the effects of training in general (e.g., Bernardin & Pence, 1980; Borman, 1979; Latham et al., 1975), without giving consideration to potential differences due to specific demands of the rating task itself. It has only recently been suggested, for example, that different rating formats may place different emphasis on the cognitive tasks required by the rater (Murphy, Garcia, Kerkar, Martin, & Balzer, 1982). Implicit in this suggestion is the notion that different training strategies might be necessary dependent on the format used. However, even beyond looking for general format x training interactions, the present results indicate that useful information might be available through further analysis of training effectiveness within particular formats. On a common sense level, just as certain ideas and/or concepts are easier to communicate

than others, it may be that certain dimensions and/or traits are easier to train than others. If, indeed, this proves to be true, then assessment of the particular characteristics associated with more easily rated dimensions and/or traits should prove valuable for both rating scale development as well as rater training efforts.

The present study also supports recent assertions that the prevailing error/accuracy negative covariation assumption may not be valid. Although some significant relationships were found between error and accuracy (e.g., between HALOCORR and the two accuracy measures and between DIST and leniency), it must be remembered that these two error measures were based on deviations from the true scores. However, even given the fact that these particular measures were derived from the true scores, their relationship with accuracy was relatively low (average $\underline{r}$ = .22), revealing only about 5 percent of shared variance. This result is comparable to recent calculations by Cooper (1981), who showed error and accuracy to share a median of only 8 percent of the variance. Further, and perhaps less optimistic with respect to our present means for assessing "errors" is that most previous research has calculated errors either without the benefit of true scores (e.g., Bernardin & Pence, 1980) or without using the true scores that were available (e.g., Borman, 1975, 1979). Calculations made similar to those researchers in the present research (i.e., the SD halo measure) revealed no relationship between error and accuracy.

Taken as a whole, there is enough evidence to suggest that a serious reevaluation of our present means for defining and measuring rating "errors" might be warranted. As alluded to in the introduction,

serious consideration should be given to the fact that highly intercorrelated dimensions and/or negatively skewed distributions, for example, may be accurate reflections of reality rather than indications of halo and leniency. Hence, researchers in the fields of Industrial Psychology and Organizational Behavior may have to reassess the assumptions that they presently embrace concerning the levels of performance that will be evidenced by a particular individual as well as among individuals within a group.

Limitations and Directions for Future Research

On a practical level, the results presented here indicate that the concern of training ought to be expanded from its exclusive concentration on rating errors to include components that are more directly focused on increasing the accuracy of performance evaluations. Similar sentiments have been echoed by a number of researchers (Borman, 1972; Ilgen & Feldman, 1983) in their contention that further advancement in the area of rater training is unlikely without the appropriate attention to a process-centered view of performance appraisal that considers the information processing functions of information gathering, storage, recall, and integration. The present study was primarily concerned with the information gathering and storage components of this process in terms of providing trainees with specific, job-relevant categories for observing and evaluating ratee performance and further developing these categories by focusing on various effectiveness levels within them. However, this research is only a first step towards attempting to increase the accuracy of raters'

evaluations. Further, there are several potential limitations to this study which indicate that caution should be exercised in drawing any definitive conclusions based on these data.

First, undergraduate students and not managers were used as raters and consequently, the results can only tentatively be generalized to a true manager/supervisor population. However, recall from the Method section that employment decisions made by students in laboratory settings have been shown to be similar to those made by professional interviewers (Bernstein, et al., 1975; Schmitt, 1976). Further, the issues addressed in the present study concerned questions of how humans process and evaluate stimuli in their environments. There is no indication from the cognitive psychology literature that this process is appreciably different for students versus "real world" appraisers of employee performance. What might be appreciably different, though, are the implicit category systems that managers/supervisors have developed versus those of the students. As previously mentioned, the categories that one uses are a function of education and experience. It thus seems logical that the category systems for assessing subordinates already in use by more experienced managers would be more well-defined than those used by a relatively inexperienced student group. Hence, convincing experienced individuals to accept a newly imposed category system might require somewhat different strategies than those employed here. Similar to many OD interventions, for example, part of the training program may have to be geared toward assessing the categories already in use by trainees and convincing the "owners" of inappropriate ones that their present means for evaluating employees is somehow inadequate (i.e., a

process analogous to "unfreezing"). Perhaps only then can acceptance and use of a newly imposed category system ensue (i.e., change and "refreezing"). It is worthwhile to note, however, that approximately half of the present subjects reported having previous experience with performance appraisals. Hence, the degree to which experience may or may not necessitate changes in the training strategy suggested here can only be evaluated by future research.

Another potential limitation of this study is that the results could be attributed to the demand characteristics of the situation. It is difficult to define, however, what constitutes demand characteristics in a training study. If subjects did change their rating behaviors in accordance with the treatment presented by the experimenter, then "demand characteristics" seem inseparable from a successful training intervention. Further, it is virtually impossible that any subject could have known the true purpose of the research. The experimenter adhered, as closely as possible, to the training programs outlined in Appendices B and C, and no discussion of the study or the hypotheses was undertaken until all data collections were complete. Students were also asked not to discuss their training sessions with others in the class. Anecdotal evidence gathered by the experimenter prior to each session suggests that subjects strictly adhered to this request.

A third potential limitation concerns the fact that observations were made from videotaped rather than live persons. It is doubtful that this limitation is severe as research reviewed by Lifson (1953) has suggested that filmed performances are rated the same as live performances. Also, in light of the inherent difficulties of obtaining

true scores from live performances, potential criticisms associated with the use of videotaped ratees do not seem particularly salient, especially considering the nature of the hypotheses under investigation here.

At a somewhat higher level of abstraction, there are several aspects of the rating process in general as well as specific consequences of categorization that were not explicitly addressed in the present study. These theoretically based limitations are nevertheless important, and they also represent potentially fruitful avenues for future research. One especially relevant issue concerns some of the consequences of categorization for memory. Recent evidence seems to indicate that there may be an upper bound on the degree to which raters are able to recall which specific behaviors a given ratee has exhibited. The reason for this is that categorization is often conceptualized as a process whereby a stimulus object/person is matched to some category prototype. Furthermore, unique behaviors emitted from a particular person become more difficult to remember over time because they are colored in such a manner as to be consistent with characteristics of the prototype to which they were matched (Wyer & Srull, 1970). That is, once a person is categorized vis a vis particular behaviors and/or characteristics, the features of the category prototype(s) come to characterize the individual. Consequently, when a rater is asked to recall information for performance evaluations, some of the information will accurately describe the person in question while other information may not (Cantor & Mischel, 1977, 1979; Sentis & Burnstein, 1979; Spiro, 1977; Tsujimoto, 1978; Tsujimoto, Wilde, & Robertson, 1978; Wyer &

Srull, 1980).

Multiple categorizations are possible (Ilgen & Feldman, 1983), however, and seem dependent on one's expertise and the degree of differentiation in the observer's category system (Rosch, 1978; Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976). The present research indicates that training can potentially be used to facilitate the development of a specialized category system that, in turn, can result in more accurate performance evaluations. However, various questions remain concerning the degree to which there may be an upper bound on the accuracy that we can ever hope to achieve. It is also quite likely that other, potentially more effective training strategies can be developed to deal with such apparently problematic issues. Consideration of the prototype matching model and its implications by future researchers may prove valuable in this endeavor.

In summary, the present research needs to be replicated and extended using different raters, possibly other rating instruments as training tools, and variants of the present training procedures that more completely address the limitations of human information processing capabilities. The effects of accuracy training over time must also be evaluated. However, given that accuracy is the crucial criterion for judging the quality of performance evaluations, the results of the present study should be viewed optimistically. They suggest that advances toward effective interventions in the area of rater accuracy training are possible.

Conclusions

In conclusion, the results of this study can be summarized in the following manner. First, RAT had the effect of increasing accuracy and decreasing leniency in subjects' ratings of videotaped managers. Second, RET decreased halo error but had no effect on leniency or accuarcy. Although the combination of RET and RAT proved somewhat less accurate than RAT alone, further research is needed to verify this finding. Finally, the dimension x training interactions suggested that the effectiveness of rater training strategies can not be considered independent of the rating format and/or the rating task itself.

APPENDIX A

RATING SCALES

STRUCTURING AND CONTROLLING THE INTERVIEW

Clearly stating the purpose of the interview; maintaining control over the interview; displaying an organized and prepared approach to the interview versus not discussing the purpose of the interview and displaying a confused approach; allowing the subordinate to control the interview when inappropriate.

High Level Performance

_____ 7 • Outlines clearly the areas to be discussed and skillfully guides the discussion into those areas.

_____ 6 • Displays good preparation for the interview and effectively uses information about the subordinate to conduct a well planned interview.

Average Performance

_____ 5 • States the purpose of the interview but fails to cover some areas he intended to discuss.

_____ 4

_____ 3 • Appears prepared for the interview but at times is unable to control the interview or to guide it into areas planned for discussion.

Low Level Performance

_____ 2 • Fails to indicate the purpose of the interview and appears to be unfamiliar with the file information.

_____ 1 • Appears unprepared for the interview and is unable to control the subordinate in the interview.

ESTABLISHING AND MAINTAINING RAPPORT

Setting the appropraite climate for the interview in a warm non-threatening manner; being sensitive to the subordinate versus setting a hostile or belligerent climate; being overly friendly or familiar during the interview; displaying insensitivity toward the subordinate.

High Level Performance

_____ 7 • Draws the subordinate out by projecting sincerity and warmth during the interview.

_____ 6 • Discusses the subordinate's problems in a candid but nonthreatening and supportive way.

Average Performance

_____ 5 • Displays some sincerity and warmth toward the subordinate and indicates by his response to the subordinate and his problems that he is reasonably sensitive to the subordinate's work-related problems.

_____ 4 • Uses mechanical means to set the subordinate at ease, i.e., offers coffee.

_____ 3

Low Level Performance

_____ 2 • Projects little feeling or sensitivity toward the subordinate; makes no friendly gestures.

_____ 1 • Is confrontive and inappropriately blunt during the interview.

RESOLVING CONFLICT

Moving effectively to reduce the conflict between Valva and the subordinate; making appropriate commitments and setting realistic goals to insure conflict resolution; providing good advice to the subordinate about his relationships with Valva, subordinates, etc. versus discussing problems too bluntly or lecturing the subordinate ineffectively regarding the resolution of conflict; failing to set goals or make commitments appropriate to effective conflict resolution; providing poor advice to the subordinate about his relationship with Valva, subordinates, etc.

High Level Performance

_____ 7 • Effectively reduces conflict between the subordinate and others by making appropriate and realistic commitments to help the subordinate get along better in the department.

_____ 6 • Provides good advice about solving problems and about improving the subordinate's poor relationships with his subordinates, Valva, etc.

Average Performance

_____ 5 • Puts forth some effort to reduce conflict between the subordinate and others but usually does not commit himself to helping with this conflict resolution.

_____ 4 • Tends to smooth over problems and provide reasonably good advice to the subordinate about conflict situations.

_____ 3

Low Level Performance

_____ 2 • Lectures ineffectively or delivers inappropriate ultimatums to the subordinate about improving his relationships with others or about changing his "attitude" toward people or problems.

_____ 1 • Fails to make commitments to help the subordinate resolve problems or provides poor advice to the subordinate about his relationships with Valva, subordinate's etc.

DEVELOPING THE SUBORDINATE

Offering to help the subordinate develop professionally; displaying interest in the subordinate's professional goals; specifying developmental needs and recommending sound developmental actions versus not offering to aid in the subordinate's professional development; displaying little or no interest in the subordinate's professional growth; failing to make developmental suggestions or providing poor advice regarding the subordinate's professional development.

High Level Performance

_____ 7 • Displays considerable interest in the subordinate's professional development and provides appropriate, high quality, developmental suggestions.

_____ 6 • Makes commitments to help professionally in the subordinate's development.

Average Performance

_____ 5 • Provides general developmental suggestions but usually fails to make a personal commitment to aid in the subordinate's professional development.

_____ 4

_____ 3 • Shows moderate interest in the subordinate's development; may direct the subordinate to seek developmental suggestions elsewhere.

Low Level Performance

_____ 2 • Expresses little or no interest in the subordinate's professional development.

_____ 1 • Fails to offer developmental suggestions or provides poor advice regarding the subordinate's professional growth and development.

MOTIVATING THE SUBORDINATE

Providing incentives for the subordinate to stay at GCI and to perform effectively; making commitments to motivate the subordinate to perform his job well, to remain with GCI, and to help GCI accomplish its objectives; supporting the subordinate's excellent past performance versus providing little or no incentive for the subordinate to stay at GCI and perform effectively; failing to make commitments encouraging the subordinate's top continued performance; neglecting to express support of the subordinate's excellent performance record.

High Level Performance

_____ 7 • A high level performer provides encouragement and appropriate incentives to pursuade the subordinate to stay with GCI and perform his job effectively.

_____ 6 • A high level performer uses compliments of the subordinate's technical expertise and excellent past performance to motivate the subordinate to meet the objectives of the department.

Average Performance

_____ 5 • Compliments the subordinate appropriately at times but is only moderately effective in using these compliments to encourage high performance, loyalty to GCI, etc.

_____ 4

_____ 3 • Provides some incentives for the subordinate to perform effectively at GCI, but generally makes few if any personal commitments to support the subordinate in his job.

Low Level Performance

_____ 2 • Fails to express support for the subordinate's past performance.

_____ 1 • Provides little or no incentive for the subordinate to remain at GCI.

APPENDIX B

RATER ERROR TRAINING

What follows is a step-by-step procedure for the trainer to follow when conducting RET. The double-spaced text is a detailed script of what the trainer will specifically say during the training. Other directions for the trainer appear in italics.

Today, you will be participating in an error training program that will help you learn how to appraise the job performance of others. Once we have finished the actual training program, I will be showing you videotapes of six managers conducting an interview with a problem subordinate. After we view each of these videotapes, you will be rating each manager on how well he conducted the interview. I will then collect these ratings, go over them, and during a regular class period, I will report back how well you did in making your evaluations.

In order to rate the behvaiors of others correctly, there are a few things you must know about how to avoid various common rating errors that can occur when you evaluate others. What I mean by rating error is any systematic fault in judgment that occurs when you appraise another person's performance. More precise definitions as well as specific examples of various errors will be discussed during this training session.

In order to demonstrate how rating errors can occur, we will be viewing two five minute videotapes similar to those you will be rating after the training program. You will actually rate the managers who appear in these two tapes, and we will discuss your ratings as a group. I am passing out packets of rating scales that you will use to make your ratings. Do not make your ratings of the manager until the tape has finished. Also, do not take notes until the tape is finished, because you might miss important parts of it. The manager that you are about to see on the first videotape is an example of a very good interviewer, who deals quite well with the problem subordinate.

Show videotape 1. When the tape is finished, ask trainees to put their first name on the first page of the rating scale packet. Give them approximately five minutes to make their evaluations of the manager.

Put trainees names on a flipchart while they are making their ratings. When trainees are finished, ask them to hand in their completed scales. Record the results of each person's ratings next to his/her name on the flipchart.

Begin discussing the discrepancies between ratings. Listed below are the true levels of performance for the manager on the first training tape.

STRUCTURING AND CONTROLLING THE INTERVIEW	3.31
ESTABLISHING AND MAINTAINING RAPPORT	3.69
RESOLVING CONFLICT	5.69
DEVELOPING THE SUBORDINATE	6.08
MOTIVATING THE SUBORDINATE	5.77

Do not mention these "true scores" to trainees. Use these scores only for your information to appropriately direct the discussion of various

rating errors.

First, look for trainees who committed halo error. This error will be evidenced by ratings that are consistently high or low across the seven individual rating scales. Try to identify one or more persons whose ratings follow this pattern, and ask them why they rated the manager as they did. One of the following two responses are likely to occur:

1. *Trainees may discuss one or two things the manager did that were good or bad. You can imply from this type of response that the trainees ratings were based on the one or two things mentioned. Make a note of any trainees who back up their ratings with examples of things that occurred early in the interview. These will be used later on as examples of first-impression effect.*

2. *If the trainee(s) rated the manager high across all the performance scales, s/he might alternatively say the reason was because you (the trainer) had said the manager was an effective performer.*

Now try to identify some trainee(s) whose ratings are not consistent across all the rating scales. Ask the person(s) to explain the reasoning behing their ratings. These responses will most likely include both strengths and weaknesses of the manager.

In any case, continue the discussion as follows:

What we have just witnessed is an example of one type of rating error called halo. The term "halo" implies that there is a general aura surrounding all the judgments that are made about a particular ratee. What typically happens is that the rater forms a generally favorable or unfavorable impression of the ratee, and then gives the person ratings that are consistent with this good or bad impression. Those of you who rated the manager high just because I told you he was an effective performer committed halo error. Those of you who formed a generally good or bad impression of the manager based on one or two

characteristics (and thus gave the manager all high or all low ratings) also committed halo error.

One thing that is important to remember is that people are not typically all good or all bad. Because of this, it is essential that you try <u>not</u> to form a general impression when rating others.

Ask trainees what can be done to eliminate halo error. It should be suggested that evaluations be made independently of what raters have heard from others, and that raters make a point of looking for both positives and negatives.

For those of you who did commit halo error, I want you to realize that this is a very common occurance. Most people do form general impressions of others which do influence subsequent appraisals of their behavior.

Now try to identify trainees who committed central tendency error. This error is characterized by ratings that are concentrated around the middle anchors of the rating scale (i.e., 3, 4, or 5). Ask trainees who committed central tendency error to explain the reasoning behind their ratings. After one or more rationales have been given, continue the discussion as follows:

When all the ratings are concentrated around the middle anchors on the rating scale, this is an example of what is called central tendency error. This error occurs when the rater is afraid to use the extremely good or the extremely bad anchors of the scale, even though the ratee is exhibiting excellent or poor performance.

To summarize where we are at this point, we have discussed two errors that can occur when evaluating the performance of others. These

errors were halo and central tendency.

What we are going to do now is view a second videotape of another manager interviewing the same problem subordinate. After the tape is finished, you will rate the manager, just as we did with the first videotape.

Show videotape 2. When the tape is finished, ask trainees to put their first name on the first page of the rating scale packet. Give trainees approximately five minutes to make their evaluations.

Put trainees names on the flipchart while that are making their ratings. Have participants hand in their ratings and record these next to their names on the flipchart.

Generate a discussion centering on any discrepancies among the ratings. During the discussion if any of the trainees compare the second manager to the first manager, this will allow discussion of contrast effects to begin. If no trainee compares the two managers, ask them how they thought the second manager did with respect to the first. Further, ask trainees if they had used the first manager as a comparison point when they rated the second. Once any discussion of comparisons occurs, continue the program as follows:

If any of you rated the second manager by comparing his performance to the first, you committed a contrast error. More specifically, a contrast error occurs when we evaluate a person by comparing him/her to someone we have just finished rating instead of evaluating the person on how well s/he has performed independently of others and relative to the job in question.

Ask trainees what we might do to minimize contrast effects. The suggestions that should be made (either by the trainer or trainees) should include: (1) evaluate the applicant in relation to his/her absolute level of performance and (2) decide what these absolute levels

of performance are before you begin evaluating people.

There is one final error we will discuss today, and it is concerned with different tendencies some raters have regardless of the person they are evaluating. For example, if any of you gave both managers generally high ratings, you may, in general, be rating others too leniently. On the other hand, if you gave both managers relatively bad ratings, you may, in general, be rating others too harshly or strictly. Raters who consistently give ratings that are either too high or too low across many ratees are committing leniency/severity error.

The difference between halo and leniency/severity is that halo is person specific. In other words, you have certain general impressions of each person, and you therefore rate some people high and some low. With leniency/severity, the problem lies in the fact that you consistently rate <u>all</u> ratees either too high (as in leniency) or too low (as in severity).

Look at all trainees ratings for both managers. Select one or more sets of ratings that are indicative of leniency/strictness error and use these as examples for the group. Ask trainees to generate ideas regarding how we might decrease the occurance of leniency/strictness in ratings.

After completion of the second rating exercise, complete the training program as follows:

All of you should now understand how various rating errors can distort our evaluations of others. You will now be rating six more videotapes of different managers interviewing the same problem

subordinate. We will not be discussing these ratings, but I will collect them and evaluate how well you did. The results of this exercise will then be reported back to you during a regular class session.

As you are observing the videotapes, keep in mind the rating errors we have discussed and the various ways they might be minimized. Try using these strategies as you view and rate each manager's performance.

APPENDIX C

RATER ACCURACY TRAINING

What follows is a step-by-step procedure for the trainer to follow when conducting RAT. The double spaced text is a detailed script of what the trainer will specifically say during the training. Other directions for the trainer appear in italics.

Today, you will be participating in a training program that will help you learn how to accurately appraise the job performance of others. Once we have finished the actual training program, I will be showing you videotapes of six managers conducting an interview with a problem subordinate. After we view each of these videotapes, you will be rating each manager on how well he conducted the interview. I will then collect these ratings, go over them, and report back to you during a regular class period how well you did in rating the videotapes.

In order to rate the behavior of others correctly, there are a few things that you must know about how performance appraisal systems are set-up. First of all, most jobs can be thought of as consisting of various categories or dimensions of performance. In fact, you can think of any job as a pie that can be cut or divided into various pieces. Whenever we evaluate an employee's job performance, it is very important that we rate the person in terms of important categories of performance.

The reason for this is because these pieces or categories are the crucial elements of the job. Therefore, in order to effectively evaluate how people are performing their jobs, it is essential that we rate them on these important dimensions.

As I mentioned before, today we will be rating six managers conducting an interview with a problem subordinate. In order to appraise the performance of these managers, the first thing we must do is identify the important elements of the task that we will evaluating. There are five performance dimensions that we will be using to rate these six videotaped managers.

What I am passing out to you now are the actual rating scales we will be using. You will notice that there are five scales, one corresponding to each important category of performance. What we are going to do now is to review each of these categories and what they mean.

The first category we will use to rate the manager's performance is how well s/he STRUCTURES AND CONTROLS the interview with the subordinate. A manager who does a good job with respect to this dimension will do such things as clearly state the purpose of the interview; he will maintain control over the interview; and he will be organized and prepared for the interview. A manager who does not perform well with respect to this category will <u>not</u> discuss the purpose of the interview; he will display a confused approach; and he will allow the subordinate to control the interview at inappropriate times.

Similarly go over all of the performance dimensions by giving a global definition of what constitutes effective and ineffective performance on

it.

Now that we have our seven performance dimensions and global definitions of each, the next thing I would like to do is give you more specific examples of what constitutes different levels of effective and ineffective performance for each category. As you have probably noticed, corresponding to the scale anchors are examples of what types of behaviors are considered High Level Performance, Average Performance, and Low Level Performance. What I would like to do now is to go over specific examples of behaviors corresponding to the different performance levels on each of the 5 categories. Then we will practice using these scales by rating a videotaped manager conducting an interview with a problem subordinate.

Go through each of the dimensions by giving specific examples of behavior corresponding to the seven levels of performance.

As I mentioned, what I would like to do now is give you some practice in using these rating scales. I am going to show you a five minute videotape, and when the tape is finished, you will rate the manager on the five performance dimensions. Do not take notes while the videotape is playing, because you might miss things that the manager does. As you are watching the tape, though, look for specific effective and ineffective behaviors the manager exhibits that correspond to our seven categories of performance. This will help you to remember what the manager actually did and how well he did it (that is, whether it was high, average, or low performance).

Show videotape 1. When the tape is finished, ask trainees to put their first name on each scale and then give them approximately three minutes to make their ratings.

Put trainees names on a flipchart while they are making their ratings. When they are finished, ask them to hand-in their rating for STRUCTURING AND CONTROLLING THE INTERVIEW. *Record the results on the flipchart next to each trainee's name.*

Generate a group discussion that focusses on any discrepancies among trainees. Make sure people discuss which particular manager behaviors they considered in making their rating. Use the scale anchor descriptions to evaluate the effectiveness of each behavior discussed. Also, make sure that any behaviors brought up are legitimate examples that correspond to the dimension in question.

Repeat this process for each of the other six dimensions/rating scales.

Listed below are the true levels of performance for the manager on the first training tape.

STRUCTURING AND CONTROLLING THE INTERVIEW	3.31
ESTABLISHING AND MAINTAINING RAPPORT	3.69
RESOLVING CONFLICT	5.69
DEVELOPING THE SUBORDINATE	6.08
MOTIVATING THE SUBORDINATE	5.77

Do not directly mention these "true scores" to trainees. Merely deal with each performance dimension by discussing specific behaviors and their effectiveness levels in terms of the dimension descriptions.

Tell trainees that they will now rate another videotape of a manager interviewing the same problem subordinate. Show videotape 2. Follow the exact instructions and procedure as you did on the first videotape.

The true levels of the manager's performance for the second training tape appear below:

STRUCTURING AND CONTROLLING THE INTERVIEW	2.79
ESTABLISHING AND MAINTAINING RAPPORT	1.50
RESOLVING CONFLICT	2.07
DEVELOPING THE SUBORDINATE	2.71
MOTIVATING THE SUBORDINATE	2.29

After completion of the second rating exercise, summarize and end the training program as follows:

All of you should now understand how to use these rating scales to evaluate the performance of a manager who is interviewing a problem subordinate. You will now be rating six more videotapes of different managers conducting the same interview. We will not be discussing these ratings, but I will collect them and evaluate how well you did. The results of this exercise will then be reported back to you during a regular class session.

As you are observing the videotapes, keep in mind the seven categories you will be rating the managers on. As we did during the practice sessions, look for specific behaviors that will help you identify which level of performance the manager is exhibiting. Also, use the anchors that appear on the rating scales themselves to help you justify your final rating decision.

REFERENCE NOTES

1. Wherry, R. J. The control of bias in rating: A theory of rating. (Personnel Research Board Rep. 922). Washington, D.C.: Department of the Army, Personnel Research Section, February, 1952.

2. Bernardin, H. J., & Boetcher, R. The effects of rater training and cognitive complexity on psychometric error in rating. Paper presented at the annual meeting of the American Psychological Association, San Francisco, 1978.

3. Phillips, J. S., & Lord, R. G. Leadership prototypes: Effects on memory for leadership behavior. Manuscript in preparation, 1982.

REFERENCES

Abelson, R. P. Script processing in attitude formation and decision making. In J. S. Carrol & J. W. Payne (Eds.), Cognition and Social Behavior. Hillsdale, N. J.:Erlbaum, 1976.

Averbach, E., & Coriell, A. S. Short term memory in vision. Bell Systems Technical Journal, 1961, 40, 309-328.

Berman, D. S., & Kenny, D. A. Correlation bias: Not gone and not to be forgotton. Journal of Personality and Social Psuchology, 1977, 35, 882-887.

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

Bernardin, H. J. Behavioral expectation scales vs. summated scales: A fair comparison. Journal of Applied Psychology, 1978, 63, 125-131.

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. A recomparison of behavioral expectation scales to summated scales. Journal of Applied Psychology, 1976, 61, 564-570.

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.

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

Bernstein, V., Hakel, M. D., & Harlan, A. The college student as an interviewer: A threat to generalizability? Journal of Applied Psychology, 1975, 60, 266-268.

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, 233-252.

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

Bousfield, W. A. The occurance of clustering in the recall of randomly arranged associates. Journal of General Psychology, 1953, 49, 229-240.

Bower, G. H., Black, J. B., & Turner, J. T. Scripts in text comprehension and memory. Cognitive Psychology, 1979, 11, 177-220.

Brown, E. M. Influence of training, method, and relationship on the halo effect. Journal of Applied Psychology, 1968, 52, 195-199.

Bruner, J. S. On perceptual readiness. Psychological Review, 1957, 64, 123-152.

Bruner, J. S. Social psychology and perception. In E. E. Maccoby, T. M. Newcomb, & E. L. Hartley (Eds.), Readings in social psychology. New York: Holt, Rinehart, & Winston, 1958.

Cantor, N., & Mischel, W. Traits as prototypes: Effects on recognition memory. Journal of Personality and Social Psuchology, 1977, 35, 38-48.

Cantor, N., & Mischel, W. Prototypes in person perception. In L. Berkowitz (Ed.), Advances in experimental social psychology, (Vol. 12). New York: Academic Press, 1979.

Cohen, C. E. Cognitive basis of stereotyping. Paper presented at the Annual Meeting of the American Psychological Association, San Francisco, 1977.

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.

DeCotiis, T. A. An analysis of the external validity and applied relevance of three rating formats. Organizational Behavior and Human Performance, 1977 19, 247-266.

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

Dunnette, M. D., & Borman, W. C. Personnel selection and classification systems. Annual Review of Psychology, 1979, 30, 477-525.

Erikson, C. W., & Collins, J. F. Temporal course of selective attention. Journal of Experimental Psychology, 1969, 80, 254-261.

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

Feldman, J. H., & Hilterman, R. J. Stereotype attribution revisited: The role of stimulus characteristics, racial attitude, and cognitive differentiation. Journal of Personality and Social Psuchology, 1975, 31, 1177-1188.

Frederiksen, C. H. Representing logical and semantic structure of knowledge acquired from discourse. Cognitive Psychology, 1975, 7, 371-458.

Glass, A. L., Holyoak, K. J., & Santa, J. L. Cognition. Reading, Mass.: Addison-Wesley Publishing Co., 1979.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1954.

Hastorf, A. N., Schneider, D. J., & Polefka, J. Person perception. Reading, Mass.: Addison-Wesley Publishing Co., 1970.

Ilgen, D. R., & Feldman, J. M. Performance appraisal: A process approach. In B. M. Staw (Ed.), Research in organizational behavior, (Vol. 2). Greenwich, Conn.: JAI Press Inc., 1983.

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

Jecker, J. O., Maccoby, N., & Breitrose, H. S. Improving accuracy in cues of comprehension. Psychology in the Schools, 1975, 24, 653-669.

Kane, J. S., & Lawler, E. E. Methods of peer assessment. Psychological Bulletin, 1978, 85, 555-586.

Kelly, G. A. A theory of personality: The psychology of personal constructs. New York: Norton, 1955.

King, L., Hunter, J., & Schmidt, F. Halo in a multidimensional forced-choice performance evaluation scale. Journal of Applied Psychology, 1980, 65, 507-516.

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

Langdale, J. A., & Weitz, J. Estimating the influence of job information on interviewer agreement. Journal of Applied Psychology, 1973, 57, 23-27.

Langer, E. J. Rethinking the role of thought in social interaction. In J. H. Harvey, W. J. Ickes, & R. F. Kidd (Eds.), New directions in

attribution research (Vol. 2). Hillsdale, N. J.: Erlbaum, 1978.

Latham, G. P., & Wexley, K. N. Behavioral observation scales for performance appraisal purposes. Personnel Psychology, 1977, 30, 255-268.

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

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

Lawrence, D. M. Two studies of visual search for word targets for controlled rates of presentation. Perception and Psychophysics, 1971, 10, 85-89.

Levine, J., & Butler, J. Lecture vs. group decision in changing behavior. Journal of Applied Psychology, 1952, 36, 29-33.

Lifson, K. A. Errors in time-study judgments of industrial work-pace. Psychological Monographs, 1953, 67 (5, White No. 355).

Lord, R. G., Binning, J. F., Rush, M. C., & Thomas, J. C. The effect of performance cues and leader behavior on questionnaire ratings of leadership behavior. Organizational Behavior and Human Performance, 1978, 21, 27-39.

Lord, R. G., Foti, R.J., & Phillips, J. S. A theory of leadership categorization. In D. L. Hunt, R. J. Sedaran, & C. L. Shriesheim (Eds.), Leadership: Beyond establishment views. Carbondale: McGraw-Hill, 1982.

Marcus, H. Self-schemata and processing information about the self. Journal of Personality and Social Psuchology, 1977, 35, 63-78.

McArthur, L. Z., & Post, D. L. Figural emphasis and person perception. Journal of Experimental Social Psychology, 1977, 13, 520-535.

Minsky, M. A. A framework for representing knowledge. In P. H. Winston (Ed.), The psychology of computer vision. New York: McGraw-Hill, 1975.

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

Naylor, J. C., & Wherry, R. J. The use of simulated stimuli and the "JAN" technique to capture and cluster the policy of raters. Educational and Psychological Measurement, 1965, 25, 964-986.

Nisbett, R. E., & Wilson, T. D. The halo effect: Evidence for unconscious alteration of judgments. Journal of Personality and Social Psuchology, 1977, 35, 250-256.

O'Brien, R. Robust techniques for testing heterogeneity of variance in factorial designs. Psychometricia, 1978, 43, 327-342.

Picek, J. S., Sherman, S. J., & Shiffrin, R. M. Cognitive organization and coding of social structures. Journal of Personality and Social Psuchology, 1975, 31, 758-768.

Potts, G. R. Information processing strategies used in the encoding of linear orderings. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 727-740.

Rosch, E. Principles of categorization. In E. Rosch (Ed.), Cognition and categorization. Hillsdale, N. J.: Erlbaum, 1978.

Rosch, E., Mervis, C. G., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. Basic objects in natural categories. Cognitive psychology , 1976, 8, 382-439.

Rumelhart, D. E. Notes on schemas for stories. In D. G. Bobrow & A. Collins (Eds.), Representation and understanding: Studies in cognitive science. New York: Academic Press, 1975.

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

Schmitt, N. Social and situational determinants of interview decisions: Implications for the employment interview. Personnel Psychology, 1976, 29, 79-101.

Schneider, W., & Shiffrin, R. M. Controlled and automatic information processing: I. Detection, search, and attention. Psychological Review, 1977, 84, 1-66.

Schwab, D. P., Heneman, H., & DeCotiis, T. Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 1975, 28, 549-562.

Sentis, K. P., & Burnstein, E. Remembering schema-consistent information: Effects of a balance schema on recognition memory. Journal of Personality and Social Psuchology, 1979, 37, 2200-2211.

Shank, R., & Abelson, R. P. Scripts, plans, goals, and understanding. Hillsdale, N. J.: Erlbaum, 1977.

Shiffrin, R. M., & Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 1977, 84,

127-190.

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

Spiro, R. J. Remembering information from text: The "state of schema" approach. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), Schooling and the acquisition of knowledge. Hillsdale, N. J.: Erlbaum, 1977.

Spool, M. D. Training programs for observers of behavior: A review. Personnel Psychology, 1978, 31, 853-885.

Sulin, R. A., & Dooling, D. J. Intrusion of thematic ideas in retention of prose. Journal of Experimental Psychology, 1974, 103, 255-262.

Swann, W. B., & Snyder, M. On translating beliefs into actions: Theories of ability and their application in an instructional setting. Journal of Personality and Social Psuchology, 1980, 38, 879-888.

Taylor, E. K., & Hastman, R. Relation of format and administration to the characteristics of graphic rating scales. Personnel Psychology, 1956, 9, 181-206.

Taylor, S. E., & Crocker, J. Scematic bases of social information processing. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), Social cognition; The Ontario symposium on personality and social psychology. Hillsdale, N. J.: Erlbaum, 1981.

Taylor, S. E., & Fiske, S. T. Salience, attention, and attribution: Top of the head phenomena. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 11). New York: Academic Press, 1978.

Tesser, A. Self-generated attitude change. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 11). New York: Academic Press, 1978.

Thorndyke, P. W. Cognitive structures in comprehension and memory of narrative discourse. Cognitive Psychology, 1977, 9, 77-110.

Treisman, A. M., & Geffen, G. Selective attention: perception or response? Quarterly Journal of Experimental Psychology, 1967, 19, 1-17.

Treisman, A. M., & Riley, J. G. A. Is selective attention selective perception or selective response? Journal of Experimental Psychology, 1969, 79, 27-34.

Triandis, H. C. Cultural influence upon cognitive processes. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 1). New York: Academic Press: 1964.

Tsujimoto, R. N. Memory bias toward normative and novel trait prototypes. Journal of Personality and Social Psuchology, 1978, 36, 1391-1401.

Tsujimoto, R. N., Wilde, J., & Robertson, D. R. Distorted memory for exemplars of a social structure: Evidence for schematic memory processes. Journal of Personality and Social Psuchology, 1978, 36, 1402-1414.

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

Wahler, R. G., & Leske, G. Accurate and inaccurate observer summary reports. Journal of Nervous and Mental Disease, 1973, 156, 386-394.

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)

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

Weiner, Y., & Schneiderman, M. L. Use of job information as a criterion in employment decisions of interviewers. Journal of Applied Psychology, 1974, 59, 699-706.

Winer, B. J. Statistical principles in experimental design, 2nd ed. New York: McGraw-Hill, 1971.

Woll, S., & Yopp, H. The role of context and inference in the comprehension of social action. Journal of Experimental Social Psychology, 1978, 14, 351-362.

Wyer, R. S., Jr., & Srull, T. K. Category accessibility: Some theoretical and empirical issues concerning the processing of social stimulus information. In E. T. Higging, C. P. Herman, & M. P. Zanna (Eds.), Social cognition: The Ontario symposium on personality and social psychology. Hillsdale, N. J.: Erlbaum, 1980.

Zandy, J., & Gerard, H. B. Attributed intentions and information selectivity. Journal of Experimental Social Psychology, 1974, 10, 34-52.