THE EFFECTS OF FRAME-OF-REFERENCE AND RATER ERROR TRAINING ON THE ACCURACY OF PERFORMANCE APPRAISALS: UTILIZING AN APTITUDE-TREATMENT APPROACH

by

Dean T. Stamoulis

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

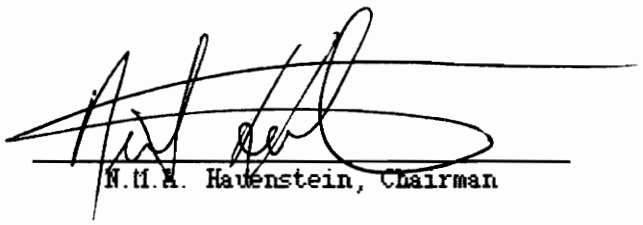
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

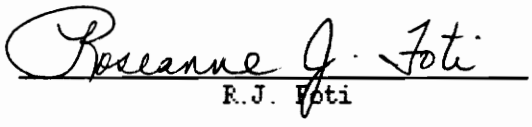
in

Psychology

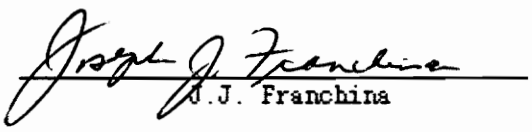
APPROVED:



N.M.A. Hauenstein, Chairman



R.J. Foti



J.J. Franchina

December, 1990

Blacksburg, Virginia

THE EFFECTS OF FRAME-OF-REFERENCE AND RATER ERROR TRAINING ON THE ACCURACY OF PERFORMANCE APPRAISALS: UTILIZING AN APTITUDE-TREATMENT APPROACH

by

Dean T. Stamoulis

Committee Chairman: Neil M. A. Hauenstein

Department of Psychology

(ABSTRACT)

Prior research has shown that frame-of-reference training increases the accuracy of performance appraisals more than rater error training (e.g. Hedge & Kavanaugh, 1988). Frame-of-reference training facilitates the learning of accurate performance standards (e.g. Athey & McIntyre, 1987), while rater error training results in the introduction of biased response sets (e.g. Bernardin & Pence, 1980). Bernardin and Buckley (1981) recommended that individuals who possessed an idiosyncratic rating style or aptitude would benefit especially from frame-of-reference training. However, no research to date has investigated the interaction of rating style and frame-of-reference training effects. The hypothesis of the present study was that rating accuracy and reliability would improve for idiosyncratic raters in frame-of-reference training, while the effective rating style of normative raters would not change. Further, rater error training should impair normative raters' accuracy and reliability, while it should not affect the ineffective rating style of idiosyncratic raters. However, the results of this study failed to show the rating aptitude-training-time interaction with accuracy. Some support was found for a rating aptitude-training-time interaction with reliability as a result of rater error training. This study replicated previous findings that frame-

of-reference training increased rating accuracy and reliability. Frame-of-reference training improved the Cronbach (1955) measures of differential elevation, stereotype accuracy, and differential accuracy.

Acknowledgements

Many people contributed their time and effort to this project. Most importantly, I would like to extend my deepest gratitude to Neil Hauenstein. His guidance and support over the past two years were crucial to the completion of this work. Second, I would like to thank Joe Franchina and Roseanne Foti. Their input improved the quality of this research dramatically. My sincere appreciation is also extended to Kelly O'Reilly, Eric Martin, and Jeanie Smith. These three performed suberbly as trainers in the rater training study. Thanks also go to three other undergraduate assistants: Darin Jacks, Andrea Conigliaro, and Shirl Innis.

Many friends supported me through the two years of this thesis project. I am indebted to Jean Markley, whose patience and support were unsurpassed. In addition, I must thank Lance Becker, Monnie Bittle, Trina Bogle, Bob Brill, Marta Carter, Stuart Greenberg, Donna Vaught, and Steve Walker. All have helped me immensely as I have enjoyed their company both in and outside the realm of this thesis.

A very special measure of gratitude also goes to members of my family--Elena Stamoulis, Mark Stamoulis, and especially to two very good friends, my parents Thomas and Chrysanthe Stamoulis. Thank you both very much for everything you have given me.

Table of Contents

Introduction

Performance appraisal systems occupy a fairly ubiquitous and influential role in most major organizations. Performance ratings often influence employee selection, compensation, and promotion. However, the actual utility of the rating process is limited severely by the subjective nature of evaluating performance. Appraisal systems are susceptible to biases that originate from a myriad of personal, contextual, and psychometric sources (Borman, 1977; Cooper, 1981; DeNisi, Cafferty, & Meglino, 1984; Landy & Farr, 1980). Such biases may reduce the accuracy of performance appraisals (e.g. DeNisi et al., 1984).

A large portion of the performance appraisal literature has been concerned with reducing bias in ratings. Research efforts have focused on two major methods of bias reduction: 1) making the rating instrument itself less prone to bias; and 2) training raters so that they may avoid bias in their appraisals. Since variations in instrument format have had a minimal impact on rating accuracy (e.g. Landy & Farr, 1980), more work has focused on developing training programs which directly reduce specific rater biases. It was assumed that such trainings would increase rating accuracy.

Many studies have shown that training can reduce common rating errors such as halo and leniency (e.g. Bernardin, 1978; Bernardin & Walter, 1977; Borman, 1975; Ivancevich, 1979; Latham, Wexley, & Purcell, 1975). Specifically, rater error training involved educating trainees about major rating errors like halo and leniency, and then instructing those trainees to reduce the incidence of the rating errors under focus. Halo is usually defined as a high intercorrelation amongst ratings across dimensions for a particular ratee. Leniency refers to a rater's overall tendency to give uncritical, favorable ratings to a ratee. Until recently, however, only three rater error training studies had used accuracy

specifically as a dependent measure (cf. Bernardin & Buckley, 1981). Interestingly, results from that trio of studies (Bernardin & Pence, 1980; Borman, 1975; 1979) were contrary to the assumption that decreased bias caused an increase in accuracy. Accuracy was not improved by such training programs, and even decreased in certain instances (Bernardin & Pence, 1980).

Rater error training seems to facilitate more than the learning of basic psychometric concepts about rater biases. Training on psychometric errors can foster the acquisition of an alternative response set in raters that results in lower accuracy. In classical test theory terms, reducing these biases may in fact be reducing substantial portions of true systematic variance in performance ratings. In trainings on leniency and halo error, the implication is that raters learn a response set of low mean ratings and low intercorrelations across dimensions, respectively (Bernardin & Buckley, 1981).

Also, experimental demand characteristics in these experimental settings are such that subjects may attend to subtle cues in order to be consistent with experimenter expectations. Bernardin and Buckley (1981) concluded that demand effects may be significant especially in the rater error treatment setting. The authors stated that the purpose of rater error training is made quite salient to the experimental subject. As the Bernardin and Pence (1980) data indicates, the new response set which is acquired does not result in more accurate and valid ratings, however.

Improvement in the accuracy of performance ratings may be dependent upon advancing a more comprehensive understanding of the appraisal process. From this perspective, training could address the systematic aspects of performance appraisal which contribute to bias (DeNisi et al., 1984). Concomitant with the findings of Bernardin and Pence (1980), concepts of social cognition began to be

applied to performance appraisal in the 1980's. Several cognitive models have been proposed and now represent the major efforts of theory building in performance appraisal (e.g. Cooper, 1981; DeNisi et al., 1984; Feldman, 1981; Ilgen & Feldman, 1983). These social cognition models attempt to describe the method by which a rater collects, encodes, stores, and retrieves information from memory. This approach also is concerned with the ways that the rater distributes weights and integrates information to form an evaluation. Performance appraisal, on the whole, is viewed as an exercise in social perception which is embedded in an organizational context and involves the utilization of both formal and implicit judgement (DeNisi et al., 1984).

To this point in time, the social cognitive approach to performance appraisal has not resulted in many innovations of practical utility. This perspective's largest contribution to date has been the introduction of an alternative rater training strategy (Hauenstein & Foti, 1989). Bernardin and Buckley (1981) proposed that one method of increasing observational skills is to establish a common frame of reference for observing and evaluating behavior. Raters are trained to adopt the same, organizationally-derived evaluative standard as the reference for judging ratee job performance. Frame-of-reference training specifically includes a discussion which focuses on defining rating dimensions, identifying work behaviors that characterize various levels of performance effectiveness within each dimension, practicing the rating of work behaviors, and feedback about the behaviors that significantly led to those ratings. Several studies have shown that frame-of-reference training facilitates more accurate and reliable performance evaluations (Athey & McIntyre, 1987; Hedge & Kavanaugh, 1988; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984, 1986).

Frame-of-reference training has several advantages over the more traditional rater error training. Rater error training results in the acquisition of a specific response set which does not necessarily improve rating accuracy. Frame-of-reference training fosters the learning of a job performance-related orientation upon which one bases performance ratings. Thus, frame-of-reference training is directly oriented towards accuracy in the appraisal process. In frame-of-reference training, raters are informed about the correct ratings for specific job behaviors, and about the rationale for each rating (Bernardin & Buckley, 1981). In most rater error training studies, participants are given information about the major psychometric rating errors and biases. They do not receive information that will aid directly in improving their rating accuracy. Finally, in frame-of-reference training, all raters are given the same information based upon normative data originating from the organization itself. The standardization of the training across raters results in more reliable ratings. Since the training is organization-specific, the training increases rating accuracy. Rater error training does not emphasize the organization in which the raters and ratees are embedded.

While frame-of-reference training may be viewed as standardized, it was intended to be so for a select population possessing a specific aptitude. Early in Bernardin and Buckley's (1981, p. 209) proposal of frame-of-reference training, they advised that:

> "in order to acquire this common frame of reference, raters with idiosyncratic standards of work performance (ie. those whose standards are not in accord with organizational norms) should be identified."

Training should focus on raters with idiosyncratic standards so that their perceptions can be brought into closer congruence with the organization. Frame-of-reference training is specifically designed to accomplish this (Bernardin & Buckley, 1981). This issue, a major part of the original proposition of frame-of-reference training, has been ignored generally in the pertinent research literature (Hauenstein & Foti, 1989).

Borman (1987) assessed the content of raters' cognitive categories. He found that the categories representing job performance possessed both similarities and differences across raters. From the perspective of Bernardin and Buckley (1981), the similarities in category content are reflective of normative data, while category content differences are indicative of idiosyncratic rater characteristics.

Hauenstein and Alexander (in press) concluded that raters with idiosyncratic frames of reference process performance information in a different manner than raters possessing a shared frame of reference. They concluded that idiosyncratic raters may be discrepant from the normative frame of reference in two ways: 1) these raters may systematically overestimate or underestimate the performance levels of important behaviors comprising the frame of reference, 2) idiosyncratic raters may be insensitive to covariation in work behaviors (Hauenstein & Foti, 1989).

The operational definitions of these two biases are identical to threshold and sensitivity analyses. These two analyses components were developed in testing the inferential accuracy model of Jackson (1972), which describes the manner in which people can accurately evaluate others based on a limited amount of information. The basic tenet of the Jackson model is that the accurate perception of others depends on the observer's implicit performance theories and

the willingness to make inferences based on the perceived relations within a particular person performance category. Nathan and Alexander (1985) concluded that this model has important ramifications for the understanding of the performance appraisal process. Since performance appraisals are by nature inferential, the accuracy of performance ratings will not only be a function of the observed performance, but also of the raters' thresholds for making performance inferences and the raters' sensitivity to the normative relationships among behaviors. Each rater uses his/her implicit theory to aid in the integration of behaviors into judgments and to determine the information stored in memory (cf. Hauenstein & Alexander, in press). The implicit theories which workers possess about their occupations are important because the theories provide a framework for interpretation of the task environment (Borman, 1987). Hauenstein and Foti (1989) posited that frame-of-reference training is a specific application of this inferential accuracy perspective to rater training and performance appraisal.

Bernardin and Buckley (1981) proposed that frame-of-reference training should include the identification of raters with idiosyncratic frames of reference because they should benefit the most from such training. Hauenstein and Foti (1989) stated that high sensitivity raters with appropriate thresholds are unlikely to derive much benefit from frame-of-reference training since the information conveyed during training would in effect be redundant. Thus, when training the whole rater population, frame-of-reference training should exhibit an aptitude-treatment interaction (cf. Hauenstein & Foti, 1989). Idiosyncratic raters should display improvement from frame-of-reference training, while raters who already possess the appropriate frame of reference should show little change in the quality of their ratings.

To date, no test of such an aptitude-treatment interaction has been conducted. Bernardin (1979) investigated frame-of-reference training with an idiosyncratic group of raters, but he did not train and evaluate a normative group of raters. Hauenstein and Foti (1989) suggest that testing for an aptitude-treatment interaction is important for two reasons. First, a more accurate assessment of the effects associated with frame-of-reference training would be obtained. The present state of frame of reference research using a general rater population may underestimate the benefits derived by idiosyncratic raters. Second, if all raters in the population receive equal benefit from training, then the theoretical explanation by Bernardin and Buckley (1981) of why frame-of-reference training is effective may be inaccurate.

In addition to investigating the aptitude-treatment interaction in conjunction with frame-of-reference training, the present study will assess the relation of idiosyncratic/ normative aptitude to rater error training. Individual differences in rating aptitude may more parsimoniously explain why frame-of-reference training increases rating accuracy and why rater error training tends to decrease the accuracy of ratings. For raters with idiosyncratic biases, rater error training may be unable to ameliorate the inaccurate response style. With normative raters, rater error training is expected to introduce biases and alternative response sets which decrease rating accuracy.

Literature Review

Rater Error Training and Rating Accuracy

Rater error training was developed under the preconception that accuracy in observing and evaluating performance can be improved by training raters to minimize rating errors. It was assumed that more accurate ratings possessed less psychometric error (Bernardin & Pence, 1980). Therefore, studies of rater error training were directed at common rating errors such as halo and leniency. Halo error is the exaggerated degree of intercorrelation amongst ratings across dimensions. Leniency error is the exaggerated incidence of favorable ratings. However, the specific operational definitions of these two rating errors has been quite variable across studies (Sulsky & Balzer, 1988).

Early performance appraisal research reflected this focus on rating biases. The zeitgeist in U.S. psychology at the early part of this century was focused primarily on the study of overt behavior. When researchers like Thorndike (1920) began to write about halo error, training interventions became oriented towards reducing biases by addressing negative aspects of rating behavior. The focus was not on the specific processes which led to those negative aspects of rating behavior.

Bittner (1948) recognized the utility of rater error training. The training consisted of a discussion focused on the operational definitions of halo and leniency. Then, trainees were simply told to avoid making those errors as they were defined. Field studies throughout the 1950's, such as Levine and Butler (1952), confirmed that such discussions in trainings did decrease the incidence of rating biases.

In the 1970's, there was a renewed surge of interest in rater training studies. Several of the studies investigated different methods of presentation

permutations in conveying rater error training content. Latham, Wexley, and Pursell (1975) used a workshop simulation approach that exceeded the earlier discussion techniques. Their approach included a modeling component in which trainees watched a supervisor rate a job candidate on videotape. In addition, the authors employed a practice and feedback component. Raters evaluated a videotaped performance segment and received feedback on their proneness in making rating errors. Trainees in the workshop condition reduced their errors to a greater extent than those trainees in a discussion-only group. Further, this three-day workshop condition showed promise in demonstrating that such trainings can result in desirable longitudinal effects (6 months).

Borman (1975) trained raters by describing halo error and presenting examples. Raters were warned not to commit halo error by justifying a general overall impression and rating an individual at the same level on all dimensions. Bernardin and Walter (1977) investigated the effects of different training procedures on psychometric error with behavioral expectation rating scales. They concluded that raters who were trained on errors prior to the observation of performance had significantly less leniency and halo error than the other groups. Ivancevich (1979) further found that these established effects of rater error training procedures could again be replicated in an industrial setting.

Bernardin (1978) attempted to assess more specifically the cognitive results of rater error training. He found a consistent relationship between individual scores on tests of various types of psychometric error and actual measurements of these errors on ratings. The author summarized that the individual who can properly identify rating errors is more careful to avoid such errors when rating performance.

> "At this point, we can only assume that the more valid ratings are those that have less psychometric error (as we measure it)" (Bernardin, 1978; p. 307).

Bernardin suggested that before the practical significance of a particular training program can be measured, the relative validities of the performance ratings must be estimated.

Before 1981, however, only three published studies involving rater error training used validity or accuracy as a dependent measure (Bernardin & Pence, 1980; Borman, 1975; 1979). Cooper (1981) concluded that the previous neglect of accuracy criteria in rater training was a by-product of the training focus on rating behavior and rating distributions. However, Borman (1975) used pretest and posttest measures to assess halo error and the validity of ratings. Degree of halo for a single rater was defined as the variance across dimensions of that individual's ratings. The validity of a particular trainee's ratings was estimated by correlating the mean ratings for ratees on each dimension with the corresponding pretest "true" scores assigned to ratees on that dimension (a measure of correlational accuracy). The results showed that halo error decreased significantly after training, but no clear trend could be identified with respect to pretraining versus posttraining validity (Borman, 1975). The author concluded that the validity of performance evaluations does not appear to be affected significantly by attempts to reduce rating errors like halo.

Borman (1979) specifically identified deficiencies in the earlier rater error training studies (e.g. Latham et al., 1975; Bernardin & Walter, 1977; Bernardin, 1978). First, while some of the training programs appeared to be successful at reducing certain rating errors like halo or leniency, a small

proportion of rating errors are usually assessed per individual study. Therefore, it was possible that other rating errors were persisting despite training or that the errors were being exacerbated by training. Second, the studies failed to investigate rating accuracy or validity by using valid scores against which to compare trainees' ratings. Borman concluded that this situation is unfortunate because accuracy or validity should be the "critical criterion" for judging the quality of performance ratings.

Borman (1979) utilized the format of the Latham et al. (1975) workshop training package. Borman's program consisted of observing and rating the videotaped performance of a job incumbent using the rating scales provided. Ratings made by each trainee were then discussed by the participants. The trainer then gave the ratings that illustrated the avoidance of the rating error being studied and decribed specifically the reasons for the ratings. At the end of the training the trainer discussed the rating error of concern, with a focus on examples of the error and ways of overcoming it. The differential accuracy measure (cf. Borman, 1979) provided accuracy scores for each rater on each job dimension. This differential accuracy measure largely resembles the correlational accuracy measure described earlier. Trainee ratings were correlated with true scores derived from the more experimentally sound external criteria of expert mean ratings, rather than pretest measures of the same sample as in Borman (1975).

Similar to his previous study, Borman (1979) again found that rater error training did not improve rating accuracy. The author posited that while it is intuitively appealing that decreasing the incidence of these rating errors would be accompanied by increased reliability and accuracy, the relationship between

these classes of criteria is not consistently strong. Borman concluded that instead of focusing solely on rating errors, rater training should focus on reliability and accuracy criteria.

To increase reliability and interrater agreement, Borman (1979) advised that training should stress the standardization of behavioral observation. Raters should be taught a common nomenclature for defining the context of the observed behavior. Interrater agreement should be reached regarding the relative importance of different behaviors as contributors to effective performance. To improve rating accuracy, Borman (1979) suggested that these consensus effectiveness levels for behaviors, and their respective weights in comprising performance dimensions, should be "correct" and uncontaminated by factors irrelevant to performance-related considerations.

Bernardin and Pence (1980) hypothesized that the common core of traditional rater error training programs stressed that certain rating distributions are more desirable than others. They stated that rater error training facilitates the learning of a new response set which may result in lower mean ratings (less leniency) and lower scale intercorrelations (less halo), but perhaps more importantly lower levels of accuracy. These hypotheses were supported by experimental results. In fact, rater error training did not improve accuracy and led to less accurate ratings than those obtained from a group of untrained raters (Bernardin & Pence, 1980). Hedge and Kavanaugh (1988) compared rater error training and frame-of-reference training, and reported that the measure of correlational accuracy decreased for raters trained on errors. These findings are pertinent for this study. While the rating accuracy of some trainees had been unaffected by rater error training, the level of accuracy of

some of the other raters declined perhaps as a result of learning inaccurate response sets in error training.

The weaknesses in rater error training stem from two major assumptions of that particular training perspective. First, it was generally assumed that the various operational definitions of rating biases represent bias or error. However, Cooper (1981) suggested that the assumption that job performance dimensions are uncorrelated may be false. That is, halo error may not be error at all. Furthermore, Bernardin and Pence (1980) posited that the assumption that ratee performance is normally distributed across workers may be false, given the common rates of work attrition and personnel selection systems. That is, there may be more effective workers in an organization than non-effective workers. Therefore, leniency error may not be an error at all. These rating error assumptions had catalyzed the operationalizations of halo and leniency bias for many years.

Second, it was generally assumed that decreasing the incidence of these biases would increase levels of accuracy. The empirical evidence indicates the opposite. Decreasing rating biases is not related to accuracy because the definition of these biases is not necessarily related to the veridical world. The tenets which underlie rating biases do not address accuracy directly. In addition, the operational definitions of these rating biases have been continually criticized because they have frequently represented different statistical assumptions and have taken a variety of different forms (Murphy & Balzer, 1981; Saal, Downey, & Lahey, 1980; Sulsky & Balzer, 1988). These differences have led to different conclusions and reduced comparability between studies.

Bernardin and Pence (1980) concluded that further research is needed to develop and validate new rater training programs which increase rating accuracy rather than train rater response sets that have the appearance of reducing rater error. The authors advised that a new perspective should be investigated which recognizes that performance-related dimensions of behavior may, in reality, be correlated and that ratee performances may not be distributed normally. By adopting such an approach, the elusive criterion of increased accuracy in rater training programs might be more easily achieved.

Frame of Reference Training and Rating Accuracy

Bernardin and Buckley (1981) summarized the rationale for a new rater training program. The authors believed that central emphasis should be focused upon training raters to observe and judge behavior more accurately, rather than simply providing specific illustrations of how to or how not to rate with regard to response distributions. The proposed model was similar to the reliability and accuracy perspective of Borman (1979).

Bernardin and Buckley (1981) suggested three areas of emphasis. First, to hone observational skills, they recommended the use of a formal diary-keeping system. This position specifically coincides with the approach of Borman (1979), who advised that observational abilities would be improved by standardizing the evaluation process and developing a common frame of reference for identifying effective and ineffective performance. Bernardin and Walter (1977) had effectively trained student raters to record critical incidents of instructors' behavior. The authors concluded that the diary-keeping component of the training program was solely responsible for an increase in interrater agreement. It appears that practice in the observation of critical incidents

focuses the raters' attention and allows for greater reliability in performance ratings. By sharpening the attention of raters to the specific information required, an appropriate cognitive set for rating is likely to be developed and enhanced (Pulakos, 1986).

In order to acquire this common frame of reference, Bernardin and Buckley (1981) suggested that raters whose standards were not originally in congruence with organizational norms should be identified. Attempts could then be made to lessen the discrepancy between the standards of these specific individuals and the standards of the organization. Frame-of-reference training was designed directly to increase rating accuracy for this particular sample of raters. Raters who already possess the organization's frame of reference should be unaffected by the training content, since they are being presented with standards they already have (Hauenstein & Foti, 1989).

According to Bernardin and Buckley (1981), the format or method of presentation of frame-of-reference training could be similar to the workshop approach format first outlined by Latham et al., (1975). Those persons with idiosyncratic standards first would read a job description and then discuss the duties and qualifications they believe would be necessary for job performance. Then the participants would be given three vignettes comprising critical incidents of performance on the job. Each of the training vignettes used would be empirically derived to fit an outstanding, average, and unsatisfactory label by using normative ratings ascribed by both workers and supervisors. Trainees then rate each vignette on behaviorally-based rating scales and submit justifications for each rating. The trainer would then inform the participants as to what the correct ratings should be for each vignette based on the

normative data, and what the rationale is for each rating. Then, a final discussion would focus on any disparity between "correct" and idiosyncratic ratings (Bernardin & Buckley, 1981).

Only six published studies can be considered tests of frame-of-reference training as proposed by Bernardin and Buckley (Athey & McIntyre, 1987; Bernardin & Pence, 1980; Hedge & Kavanaugh, 1988; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984; 1986). All of these studies have found evidence that frame-of-reference training improves some measure of accuracy. However, each of these studies deviates in some way from the original propositions of Bernardin and Buckley (1981).

Even though Bernardin and Pence (1980) trained raters before Bernardin and Buckley's (1981) frame-of-reference training proposal, Bernardin and Pence's training contained frame of reference principles. The 45-minute frame-of-reference training consisted of a lecture on the multidimensionalty of jobs and the importance of fair and accurate ratings. Discussion then centered on generating and defining dimensions of job performance and critical examples of high, medium, and low levels of effectiveness behaviors for each dimension. There was no practice and feedback component to the training. Experimental participants were a group of undergraduates undifferentiated by rating style. The dependent measure of accuracy was the average absolute deviation of each rater's ratings on each dimension from a set of target scores derived from an independent sample. Frame-of-reference training resulted in more accurate ratings than rater error training, but there was no difference between frame-of-reference training and a control group.

In McIntyre, Smith, and Hassett (1984), the 30-minute frame-of-reference training consisted of a lecture and discussion on the dimensions of job performance, and one practice vignette with feedback of target scores and behavioral rationales of the ratings. Unfortunately, there was only one practice videotape and no final discussion focused on questions concerning the rating feedback. Experimental subjects consisted of a group of undergraduates undifferentiated by rating style. The dependent measures of accuracy were Borman's (1977) correlational measure of accuracy and the average absolute deviation measure. Frame-of-reference training resulted in higher accuracy than rater error training or a control group, especially for the deviational measure of accuracy.

Pulakos (1984) utilized a 90-minute frame-of-reference training consisting of a lecture on the multidimensionality of jobs, a discussion of the job performance dimensions, and two practice vignettes with feedback of target scores and behavioral rationales. Pulakos used only two practice vignettes, and it was not clear whether a final discussion was included to answer questions about the feedback. Experimental participants consisted of a group of undergraduates undifferentiated by rating style. The dependent measure of accuracy was Borman's (1977) correlational accuracy measure. Frame-of-reference training resulted in higher accuracy when compared to rater error training and control groups.

Pulakos (1986) investigated the effects of congruent versus incongruent rating task/ rater training combinations. The training formats of both conditions (evaluative versus observational) were similar to Pulakos (1984). Experimental participants consisted of a group of undergraduates

undifferentiated by rating style. The dependent measures of accuracy were all four of Cronbach's (1955) deviational components of accuracy. Both frame-of-reference-types of trainings resulted in increased accuracy for their congruent task conditions, and both were more accurate than the respective control groups. This finding is significant when considering the applied implementation of frame-of-reference training.

Athey and McIntyre (1987) used a frame-of-reference training which was similar to McIntyre, et al. (1984). Experimental participants consisted of a group of undergraduates undifferentiated by rating style. Also, the dependent measures of accuracy were identical to McIntyre, et al. (1984). Frame-of-reference training improved the deviational measure of accuracy and the retention of training information as measured by a test, over a job dimension information-only group and a control group.

In Hedge and Kavavaugh (1988), a two-day frame-of-reference training consisted of a lecture focused on job performance dimensions and teaching participants to be careful observers of behavior. Job performance dimensions were also discussed, along with several rating errors. Two practice vignettes were then rated, and feedback on target scores and discussion of behavioral rationale followed. Videotapes modeling the correct use of the observational strategies were then finally presented. Although job performance dimensions seem to have been discussed, it is unclear whether behavioral examples of various effectiveness levels for the dimensions were involved. There were also only two practice vignettes involved in the training. Experimental participants consisted of a group of university supervisors undifferentiated by rating style. The dependent measure of accuracy was Borman's (1977) correlational accuracy

measure. Frame-of-reference training, along with a decision-making training, showed a limited increase in accuracy after orthogonal comparisons were performed for each job performance dimension. The authors concluded that rater error training detrimentally affected rating accuracy.

Two major components seem to be essential to frame-of-reference training. First, the training involves a discussion of job performance dimensions focusing on behavioral examples of a range of performance effectiveness. Second, the training uses a set of practice performance vignettes which display a range of job performance. Feedback should be offered via normative target scores for each dimension. Behavioral rationale for the target ratings should also be included.

The acquisition of performance standard information by means of these two components is integral to the frame of reference approach. Pulakos (1984) concluded that frame-of-reference training was more effective than rater error training on dimensions that were more explicitly defined in terms of particular effective, average, and ineffective behavioral cues. Athey and McIntyre (1987) similarly stated that the provision of rating standards and behavioral examples in frame-of-reference training was solely responsible for the improvement of rating accuracy.

Significantly, the frame-of-reference training studies ignored the differentiation of the idiosyncratic rater sample from the general population of raters that was suggested by Bernardin and Buckley (1981). Therefore, the training cannot be adequately assessed as to its impact on those individuals with idiosyncratic rating styles (the sample for which the training was specifically designed). To date, no comparison has been made between the

normative and idiosyncratic raters' change in level of accuracy as a result of frame-of-reference training. This study, however, employs a comparison of the effects of frame-of-reference training on both idiosyncratic and normative raters.

The introduction of frame-of-reference rater training into the performance appraisal literature has aided in focusing dependent variables in terms of accuracy instead of rater biases. However, there is little agreement across studies concerning the specific accuracy measures utilized. This reduces the comparability between studies. Sulsky and Balzer (1988) note that the four deviational accuracy components developed by Cronbach (1955) are based upon the true psychometric conceptualization of accuracy which includes both correlational and distance information in relation to normative target scores. In examining the six frame-of-reference studies, only Pulakos (1986) utilized the four Cronbach (1955) deviational components of accuracy. Bernardin and Pence (1980), McIntyre et al. (1984), and Athey and McIntyre (1987) used a distance index of accuracy. McIntyre et al. (1984), Pulakos (1984), Athey and McIntyre (1987), and Hedge and Kavanaugh (1988) used Borman's (1977) flawed operationalization of Cronbach's differential accuracy measure. Becker and Cardy (1986) and Sulsky and Balzer (1988) have noted that Borman's (1977) operational definition is not equivalent to Cronbach's accuracy measure. Sulsky and Balzer concluded that Borman's differential accuracy score represented a measure of correlational accuracy which did not coincide with their full psychometric definition of accuracy. This correlational measure does not qualify as an index of accuracy because it is insensitive to the distances between ratings and target scores (cf. Sulsky & Balzer, 1988). Therefore, the

assertions of most of the researchers regarding the accuracy effects of frame-of-reference training may be questionable.

The added advantage of the Cronbach (1955) accuracy components is that they include significant differential variance information. The first component, elevation, reflects overall rater variance in ratings. The second component, differential elevation, expresses ratee variance in performance ratings. The third component, stereotype accuracy, reveals dimensional variance. The fourth component, differential accuracy, represents the ratee-dimension interaction variance. These Cronbach (1955) deviational accuracy scores were used in this study. Stereotype accuracy and differential accuracy were hypothesized to reveal that frame-of-reference trainees become more accurate in approximating important aspects of the target score of a ratee. This was expected to occur since frame-of-reference training contains ratee evaluation standards on each performance dimension which should affect the dimensional and ratee-dimension information integration of performance ratings.

Measures of elevation should reflect an increase in accuracy as a result of rater error training. This is hypothesized to occur since most raters originally evaluate performance too leniently across different ratees. Because rater error trainees may have little concern about particular ratees and rating dimensions, rater error training raters' overall means should become lower and more accurate through the learning of the rater error training response set which reduces leniency.

In addition to accuracy, this study also includes rating variance or reliability criteria. Hedge and Kavanaugh (1988) observed that observation (frame of reference) and decision-making trainings both slightly increased

rating accuracy. The frame-of-reference training group also displayed an increased degree of halo, or decreased rating variance across dimensions. Frame-of-reference training probably decreased the variance of ratings through the provision of standards. This resulted in increased rating accuracy (Hedge & Kavanaugh, 1988). In contrast, rater error training fostered a reduced halo response set which increased the variance of ratings and decreased accuracy (Pulakos, 1984).

The existing frame-of-reference training studies may be plagued by two task by task experimental confounds. These confounds may limit conclusions about frame-of-reference training's effect on rating accuracy. First, in the studies of Athey and McIntyre (1987), Hedge and Kavanaugh (1988), and McIntyre, et al. (1984), practice and feedback were largely a part of the frame-of-reference training. Practice and feedback were not included in the other training groups, however. It can be concluded that a practice and feedback component is not just the domain of the frame-of-reference training approach and should not be treated as such, since earlier rater error training studies such as Latham, et al. (1975) utilized a practice and feedback format.

In a similar manner, the amount of information about the multidimensionality of jobs and the specific information about job dimensions should also be kept equal across training groups. Therefore, participants in all trainings would be drawing from the same, general job information base in observing behavior and evaluating performance.

The second possible confound is that the duration of training time differed between groups in Athey and McIntyre (1987) and in McIntyre, et al. (1984). In Athey and McIntyre (1987), the frame-of-reference training (30

minutes) appears to have been longer than rater error training (15 minutes) and no training. In McIntyre, et al. (1984), frame-of-reference training (30 minutes) was longer than both the information-only (20 minutes) and no training conditions.

Clearly, rater training researchers must be concerned with balancing training format across conditions while simultaneously remaining faithful to the theoretical approach of each training (e.g. training on errors, versus accuracy training with behavioral standards). While the information conveyed in the types of rater training programs should necessarily differ, training methods of presentation should be as similar as possible when comparing and contrasting the trainings' effects. This position directed the design of this study. It allowed for a more exacting and conservative examination of the theoretical perspectives underlying each rater training.

Idiosyncratic Rating Style

As stated by Bernardin and Buckley (1981), the tendency for raters to exhibit individual differences in their personal standards of work performance is central to frame-of-reference training. The authors recommended that raters with idiosyncratic frames of reference be identified so that they might benefit from a training intervention. Training raters who already possess the organization's frame of reference would not be an efficient use of training resources (Hauenstein & Foti, 1989). Several different perspectives have recently contributed to the investigation of individual rating styles.

Borman (1983) espoused on personal construct systems and implicit personality theory, and their implications for performance rating. The author stated that the content of a rater's personal constructs related to work

behavior may affect what the rater attends to in observing persons at work. Such "folk" or implicit theories of work performance that are held by those responsible for performance appraisal may help shape judgments about how effectively employees are performing. Borman (1987) found that the content of raters' cognitive categories which represented job performance possessed both similarities and differences across the rater population. Different experimental subjects emphasized different combinations of the performance schemata identified in the study. These differences are most likely due to experiential and general cognitive ability differences. Hauenstein and Alexander (in press) also found that raters who had idiosyncratic frames of reference processed ratee job performance information differently than raters who had a normative frame of reference.

Borman (1983) concluded that differences in these personal constructs were an important source of interrater disagreement. The culmination of his suggestions concerning future research focused on assessing the impact of these individual differences in the observation of work behavior and the rating of work performance. He recommended trying to actively impose upon raters a single, common personal construct system which corresponds to the dimensional rating system they should be using to make performance judgments.

Likewise, Nathan and Alexander (1985) stated that emphasis should be placed on raters' implicit theories of performance. The authors suggested that it would be important to identify those individuals whose idiosyncratic implicit theories are different from the normative implicit theories held by most of the other employees in the organization. Once identified, these raters could be trained on both the correct relationship between behaviors and performance

judgments (sensitivity), and on the correct level of performance (threshold). The authors stated that raters possessing implicit theories of the target occupation that are similar to the norm in terms of sensitivity and threshold (i.e. normative raters) are likely to be good judges of performance in that occupation. Raters possessing implicit theories of the target occupation that deviate from the group average in terms of sensitivity and threshold (i.e. idiosyncratic raters) are likely to be poor judges of performance in that occupation (cf. Hauenstein & Alexander, in press).

Hauenstein and Foti (1989) and Hauenstein and Alexander (in press) tested important issues concerning the applied implementation of frame-of-reference training, including the identification of idiosyncratic raters. The authors concluded similarly to Nathan and Alexander (1985) that idiosyncratic raters can deviate from the consensus frame of reference in two ways. First, raters may be incorrect in their estimation of the performance level of behaviors comprising the frame of reference. Raters with this threshold bias are identified by comparing the deviations between each rater's performance-level ratings of job behaviors in the frame of reference and the "true" performance-level values normatively assigned to these behaviors. Second, idiosyncratic raters also may be insensitive to the covariation of behaviors representing the frame of reference. Raters possessing this sensitivity bias are identified by correlating each rater's performance-level ratings of work behaviors making up the frame of reference with the "true" values assigned to these behaviors.

Hauenstein and Foti's (1989) and Hauenstein and Alexander's (in press) operational definition for idiosyncratic rating style included both the threshold and sensitivity biases simultaneously. Normative raters were

operationally defined as possessing an absence of both biases. Controlling for intelligence differences, Hauenstein and Alexander (in press) found that highly sensitive raters with moderate thresholds (i.e. normative raters) provided more accurate judgements of performance. Insensitive raters with extreme thresholds (i.e. idiosyncratic raters) were poorer judges of performance. In addition, normative raters exhibited greater interrater reliability in comparison to idiosyncratic raters.

Summary and Hypotheses

London and Hakel (1974) and the authors of rater error and frame-of-reference training studies alike have suggested that training procedures may alter the responses of raters. While frame-of-reference training has been shown to hold the most promise in increasing the accuracy of raters, certain aspects of the theory as originally put forth by Borman (1979) and Bernardin and Buckley (1981) have not yet been fully tested. Scant attention has been given to the identification of individuals who possess idiosyncratic rating styles when rating work performance. These persons are specifically expected to benefit the most from frame-of-reference training. Individuals who already observe and evaluate job performance effectively have little need for a rater training intervention. Hauenstein and Foti (1989) have posited that such normative raters should not show a change in rating quality since the information which these individuals receive in frame-of-reference training is redundant. The present study is a test of these propositions from the frame of reference model. The rating accuracy of idiosyncratic raters should increase to a larger extent than that of normative raters as a result of frame-of-reference training.

The identification of these extreme rating aptitude groups in the rater population may also aid in the explanation and prediction of rater error training effects. Idiosyncratic raters whose standards disagree with those of the organization should show no significant change in the accuracy of their ratings as a result of their introduction to another, similarly inaccurate response set. In some studies, the accuracy of raters has actually decreased as a result of rater error training (Bernardin & Pence, 1980; Hedge & Kavanaugh, 1988). This phenomenon would be expected when normative raters, whose evaluative standards are similar to the organization, learn an inaccurate rating style. Therefore, the theoretical underpinnings of frame of reference research may also explain why rater error training usually results in decreased rater accuracy.

In the present study, undergraduate experimental subjects were identified as idiosyncratic and normative raters. They were randomly assigned to either a frame of reference training group, a rater training group, or a control group. For each training condition, each group was theoretically-oriented and received practice and feedback. To assess treatment effects, pretraining rating sessions, posttraining rating sessions, and the trainings themselves utilized stimulus videotapes of work performance. Importantly, the target scores of the rating segments were derived from a sample external to the training experiment. However, the characteristics of the target score subjects were similar to the training subjects since they originated from the same pool.

The goal of this thesis is to explain the specific effects of frame-of-reference and rater error training. It is surmised that the differences in rating accuracy and dispersion which occur as a result of these two rater

training procedures take place because of an rating aptitude-treatment interaction. The main hypotheses are:

1. The rater error training condition should increase the elevation accuracy component of both the normative and idiosyncratic groups, as compared to the frame-of-reference and control training conditions. Through the learning of a less lenient response set which does not focus on differential dimension and ratee information, the overall means of most raters should become lower and more closely approximate the grand mean component of the target score.

2.a. Frame-of-reference aptitude-treatment interaction with accuracy

As a result of frame-of-reference training, normative raters should not change in stereotype accuracy or differential accuracy because they already possess appropriate dimension and ratee-dimension standards. Idiosyncratic raters should increase in accuracy because of the learning of appropriate dimension and ratee-dimension standards that they did not originally possess.

2.b. Frame-of-reference aptitude-treatment interaction with reliability

As a result of frame-of-reference training, the interrater reliability of previously reliable normative raters should be unchanged. The interrater reliability of previously unreliable idiosyncratic raters should increase.

3.a. Rater error aptitude-treatment interaction with accuracy

As a result of rater error training, normative raters should decrease in stereotype accuracy and differential accuracy because they learn a response set which overrides the appropriate dimension and ratee-dimension standards that

they originally possessed. Idiosyncratic raters should not change in accuracy because they already possess a response set which does not include appropriate dimension and ratee-dimension standards.

3.b. Rater error aptitude-treatment interaction with reliability

As a result of rater error training (i.e. decreasing the incidence of halo error), the interrater reliability of previously reliable normative raters should decrease. The interrater reliability of previously unreliable idiosyncratic raters should be unchanged.

Method

Development of Stimulus Videotapes and Target Scores

Videotaped work performance vignettes were used as stimuli to be rated in the pretraining and posttraining phases of this experiment. In addition, these videotapes were also viewed in the practice portions of the three types of training sessions.

The rating of videotaped ratee performance has been a common component of rater training studies (e.g. Athey & McIntyre, 1987; Ivancevich, 1979; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984; 1986). Pulakos (1984) summarized the recommendations of Borman (1977) by stating that videotaped performances can enable the calculation of true (or target) scores, thereby allowing an assessment of rating accuracy. The videotapes can be carefully developed to ensure that the standardized presentation of work performance represents a variety of effectiveness levels on different rating dimensions. In addition, these recorded segments of behavior may be more similar to actual performance appraisal situations than previously-used written vignettes.

This study used new videotaped vignettes developed especially for this investigation. The chief advantage to this is the opportunity to increase the quality and realism of the performances. The videotapes in this study involved a job interview situation. The work performance of the interviewer was the focus of the subjects' performance ratings.

In the construction of the videotapes, interviewer behaviors were derived from a study conducted by Hauenstein (1987). A range of these behaviors were chosen for inclusion into the videotapes based upon their relationship to effective work performance and the ease in which the behaviors could be translated via the video medium (see Appendix A). Twelve behaviors were chosen

from each of the effective, average, and ineffective performance categories to represent a broad range of performance. To help control for any rating scale or format biases which might affect the translation of the original results to this experiment, the same 7-item interviewer rating scale from Hauenstein (1987) was utilized in the development of this study's videotapes (see Appendix B). This rating form will also be used in the proposed rater training study. The dimensions on this rating form are similar to Borman (1977) and will be described later.

Scripts were carefully prepared based upon the interviewer behaviors. The interview segments were written so that two behaviors of the same category of performance effectiveness were included. Therefore, 6 interview segments were composed for each of the effective, average, and ineffective job performance categories (18 segments in total). Care was taken to preserve realism and match behaviors which might naturally covary, and to nest these critical behaviors amidst relatively innocuous interview content. Each segment was two to three minutes in duration and contained the same two actors portraying the interviewer and interviewee. Finally, to assure quality, each of these videotaped segments was filmed at the same office set and was professionally produced by the Virginia Tech Learning Resources Center.

A pilot study was then conducted to establish the target scores of the interview segments so that rating accuracy in this study could be assessed. The 18 videotaped vignettes were presented to 193 undergraduates within randomly ordered triads of the effective, average, and ineffective performance categories. These raters originated from the same pool of subjects from which raters in this proposed study will be derived. This assures that the raters' characteristics are as similar as possible. However, the resulting target scores of the

interview segments are based on subjects who did not participate in the rater training study.

This study contrasts with some of the general performance appraisal studies in that expert raters were not used in deriving target scores. Borman (1977) suggested that if experts are given enhanced opportunities to examine videotapes of job performance, the mean rating computed over a number of expert judges provides a "true" score measure of a ratee's performance. The accuracy of performance evaluations can then be assessed by comparing subjects' ratings with average ratings provided by a group of expert raters. However, frame-of-reference training specifically views "true" scores as being embedded in a particular context. Indeed, one of the strengths of this training is that raters are given information based on normative data originating from a specific performance environment. In this study, the context of concern has been defined as undergraduates rating the job performance of an interviewer in a laboratory setting. The ultimate focus of the experimental task is the training of raters so that they may approximate a designated standard. The pilot study also involved a relatively large sample size (193 subjects). Such an increased sample size enlarged the statistical power of the mean performance ratings. The number of expert raters in rater training studies is commonly, in contrast, much lower. For example, McIntyre et al. (1984) used 6 expert raters to estimate target scores, while Borman (1979) utilized 8 experts.

Hauenstein and Alexander (in press) used the generalizability model to calculate interrater reliability. In the Cronbach, Gleser, Nanda, and Rajaratnam (1972) generalizability model of reliability, facets or variables are chosen by the investigator to create a whole universe of facets that may each account for rating variance. Each facet can be individually examined in a generalizability

study to determine whether performance ratings can be generalized across values or conditions of a particular variable (Sulsky & Balzer, 1988). The relevant formulae delineated in Cardinet, Tourneur, and Allal (1976) were utilized in this study.

Identification of Idiosyncratic and Normative Raters

The process described by Hauenstein and Foti (1989) and Hauenstein and Alexander (in press) which identifies idiosyncratic raters by means of threshold and sensitivity criteria was utilized in this study. Threshold was operationally defined as the deviation between each rater's mean response level to all items of an inferential accuracy test and the grand mean of a normative sample aggregated over all items and subjects. A threshold score of zero indicates that the rater's threshold is perfect. A large negative threshold score indicates that the rater systematically underestimates performance levels. A large positive threshold score indicates that the rater overestimates performance levels. Sensitivity will be operationally defined for each rater as the correlation between an individual rater's responses to each item on the inferential accuracy test and the item's mean across all subjects. A correlation of 1.00 indicates that the rater is perfectly sensitive to the frame of reference.

A pilot study was conducted to establish the sensitivity and threshold norms used in identifying normative and idiosyncratic raters for the trainig study. The advantage of setting cutoffs to use in the identification of normative and idiosyncratic raters is that the subsequent rater training subject selection study and rater training study could be run simultaneously. Two hundred and five subjects from the psychology experimental subjects pool responded to 100 items from Form E of the Personality Research Form (PRF) in the norming study. The PRF was the inferential accuracy or rater aptitude test used

in this study. The 100 items of the PRF were derived from ten 10-item PRF scales which were originally chosen by Hauenstein and Alexander (in press) to describe interviewers.

The criteria for establishing the sensitivity and threshold norms from this group followed the original procedures outlined by Hauenstein (1987). To identify normative raters, Hauenstein used subjects who scored in the middle 50% of threshold scores and the top 50% of sensitivity scores. Idiosyncratic raters were identified as the subjects who scored in the 50% comprising the two tails of the threshold scores, and in the bottom 50% of the sensitivity scores. Following this format, a median sensitivity correlation of .66 was found in the norming sample. In addition, the first and third quartile threshold scores in the norming sample were -.15 and .21, respectively. These threshold scores possessed a negative skew which was also found by Hauenstein (1987).

These cutoffs were then used in the subsequent selection study specifically to identify normative and idiosyncratic raters for inclusion in the rater training study. Subjects who were classified as normative raters had a sensitivity correlation higher than .66, and a threshold score between -.15 and .21. Subjects who were classified as idiosyncratic raters had a sensitivity correlation less than .66, and a threshold score below -.15 or above .21.

To select the normative and idiosyncratic raters for the rater training study, 309 subjects from the psychology experimental subjects pool again responded to the 100 items from the PRF. Normative and idiosyncratic raters were identified via the procedures described above. Hauenstein (1987) found that about 25% of the total sample of raters were identified as normative, while 30% possessed an idiosyncratic rating style. In this subject selection study, the results were similar. Approximately 26% of the total sample of raters were

identified as normative, while 31% possessed an idiosyncratic rating style. Subjects who chose to take part in the subsequent rater training study were randomly assigned to one of the three experimental conditions. Forty-one subjects were in the frame-of-reference training group, forty-five subjects were in the rater training group, and forty-three subjects were involved in the control training condition. Sixty-three subjects were classified as normative raters, and sixty-six subjects were classified as idiosyncratic raters.

One issue of concern is that the idiosyncratic and normative raters represented non-equivalent groups at the pretraining measurement occasion. While the use of a rater aptitude control group might be advisable, no pertinent comparison group exists. The operational definitions of idiosyncratic and normative raters are such that the other individuals not identified as idiosyncratic or normative cannot be placed on a relevant continuum between the two groups. To compensate for this situation, it may be noted that specific hypotheses concerning these two groups have been put forth before the commencement of the formal rater training study.

Subjects

One hundred and twenty-nine subjects originating from a pool of volunteers in Virginia Polytechnic Institute and State University's undergraduate psychology courses took part in the training experiment. Forty-five subjects were males, and eighty-four subjects were females. Gender has not been deemed a relevant variable in the rater training literature, and was not controlled for in this study. The students received extra credit points for their participation.

Experimental Design

Idiosyncratic and normative rating groups were crossed with frame-of-reference training, rater error training, and control conditions. The

measurement of performance appraisal ratings occurred at pretraining and posttraining phases. Thus, the design was of a 2 (rating style) by 3 (training type) by 2 (measurement occasion) configuration.

Treatment Conditions

Three criteria were of concern when the rater training procedures were designed. First, the trainings were closely aligned with their respective theoretical positions. Second, the trainings were designed to be as similar as possible to other studies in the pertinent research literature. Third, the trainings were also intended have as similar a format as possible. Meeting these three criteria facilitated the close examination of the effects of these rater trainings as a function of rating style. An overview of the procedures involved in each training session are included in Appendix C. The specific script or protocol for frame-of-reference training is included in Appendix D. The protocol for rater error training can be found in Appendix E. The protocol for the control training is located in Appendix F.

The frame-of-reference training in this study was designed specifically to follow the recommendations of Bernardin and Buckley (1981) and to coincide as much as possible with the existing frame-of-reference training studies. The frame-of-reference training consisted of the following components:

1. Overview. Subjects were lectured on the general aspects of performance appraisal, the multidimensionality of work, and the general need to pay close attention in observing work behaviors.

2. Job Behaviors and Rating Dimensions. A group discussion was conducted which focused on generating work behaviors required in the job of interviewer. Information describing the job of the interviewer was provided through the

presentation and discussion of the 6 content rating dimensions listed on the rating scale. The name of each dimension was written on a chalkboard.

3. Training Concepts. A brief lecture then focused on conveying frame-of-reference training concepts. Trainees were told that rater accuracy can be fostered by the rater knowing what are effective, average, and ineffective examples of job behavior within each job dimension. A group discussion was conducted to list examples of effective and ineffective behaviors in each job dimension. These examples were written on the chalkboard, next to each appropriate job dimension.

4. Examples. The group watched videotaped examples of effective, average, and ineffective performance for each of the 6 rating dimensions. With each example, the trainer indicated the correct or target rating for that performance on that dimension. The trainer also pointed out specific behaviors which led to each particular rating.

5. Practice. The trainees were shown 3 more interview segments so that they could use the frame-of-reference training concepts and practice the performance rating task. The 3 vignettes represented a continuum of performance effectiveness which were derived from the target scores of the videotaped segments. In addition, the experimental subjects were instructed to write out justifications for their ratings.

6. Feedback. The trainer displayed each trainee's ratings for each ratee and dimension on the chalkboard. The trainer also wrote the correct or target ratings on the chalkboard. Each rater was then asked in turn what their behavioral rationale was for rating each of the three vignettes. The trainees were encouraged to respond based on their written justifications for their ratings. The group then discussed how on target that particular rater's

behavioral rationale was in rating each vignette. The trainer then indicated the ratings that were similar and different from the target scores. The trainer briefly talked about the strengths and weaknesses in that rater's rationale for his or her ratings. At the end of this component, the trainer asked if there were any questions the trainees may have regarding discrepancies between their ratings and the target scores, or between their behavioral rationale for rating and the behavioral rationale that was given by the trainer.

7. Summary. A summary of frame-of-reference training was presented in a discussion format.

The rater error training procedure used in this study was as follows:

1. Overview. Subjects were lectured on the general aspects of performance appraisal, the multidimensionality of work, and the general need to pay close attention in observing work behaviors.

2. Job Behaviors and Work Dimensions. A group discussion was conducted which focused on generating work behaviors required in the job of interviewer. Information describing the job of the interviewer was provided through the presentation and discussion of the 6 content rating dimensions listed on the rating scale. The name of each dimension was written on a chalkboard.

3. Training Concepts. A brief lecture then focused on conveying rater error training concepts. Halo and leniency rating errors were defined, and trainees were told how they could avoid committing the errors by spreading their ratings across dimensions and by lowering their ratings, respectively. A group discussion then produced examples of halo and leniency in everyday life and in the work environment. These examples were written on the chalkboard.

4. Examples. The trainees watched 6 videotaped vignettes of interviewer performance. After 3 of the performance vignettes, the trainer showed the

trainees a set of ratings which exemplified halo error on the chalk board. After the other 3 vignettes, the trainer showed the trainees a set of ratings which represented leniency error on the chalk board.

5. Practice. The trainees were shown 3 more interview segments so that they could use the rater error training concepts and practice the performance rating task. The 3 vignettes represented a continuum of performance effectiveness which were derived from the target scores of the videotaped segments. In addition, the experimental subjects were instructed to write out justifications for each rating.

6. Feedback. The trainer displayed each trainee's ratings for each ratee and dimension on the chalkboard. In addition, the sets of ratings which helped to exemplify and define the two rating errors were referred to in the course of discussion. Each rater was asked in turn what their rating error rationale was for rating each of the three vignettes. The trainees were encouraged to respond based on their written justifications for their ratings. The group then discussed how on target that particular rater's rationale was in rating each vignette. The trainer briefly talked about the strengths and weaknesses in that rater's rationale for his or her ratings. At the end of this component, the trainer asked if there were any questions the trainees may have regarding discrepancies between their ratings and the rating error definitions, or between their rationale for rating and the rationale that was given by the trainer.

7. Summary. A summary of rater error training was presented in a discussion format.

The control sessions replicated the other trainings' formats and involved the elements of the training sessions which were not embedded in the specific content of the frame-of-reference or rater error training perspectives.

1. Overview. Subjects were lectured on the general aspects of performance appraisal, the multidimensionality of work, and the general need to pay close attention in observing work behaviors.

2. Job Behaviors and Work Dimensions. A group discussion was conducted which focused on generating work behaviors required in the job of interviewer. Information describing the job of the interviewer was provided through the presentation and discussion of the 6 content rating dimensions listed on the rating scale. The name of each dimension was written on a chalkboard.

3. Training Concepts. A brief lecture then focused on conveying concepts in the organizational psychology subject area of job satisfaction. A group discussion then produced examples of job satisfaction issues. These examples were written on the chalkboard.

4. Examples. The trainees watched 6 videotaped vignettes which were examples of interviewer performance.

5. Practice. The trainees were shown 3 more interview segments so that they could practice the performance rating task. The 3 vignettes represented a continuum of performance effectiveness which were derived from the target scores of the videotaped segments. In addition, the experimental subjects were instructed to write out justifications for each rating.

6. Feedback. The trainer displayed each trainee's ratings for each ratee and dimension on the chalkboard. Each rater was asked how they rated each of the three vignettes. The trainees were encouraged to respond based on their written justifications for their ratings. The group then discussed how on target that particular rater was in rating each vignette. The trainer then generally indicated the ratings level of similarity to the other trainees' ratings. At the

end of this component, the trainer answered any questions the trainees may have had.

7. Summary. A summary of the job satisfaction topic was presented in a discussion format.

The three experimental conditions consisted of 7 basic components and lasted for approximately the same duration of time. All three trainings began with a general lecture-oriented overview of the relevant subject matter. Each of the three trainings also conveyed identical information about general job behaviors and rating dimensions. The third component involved the individual training concepts central to each approach. Then, specific examples consistent with each training were discussed, and then shown via the videotaped job performance vignettes. Significantly, each of the three trainings contained identical job performance example content. The fifth component consisted of practice performance ratings on three tapes representing a broad range of performance. Similar to the examples component, the videotaped performance vignettes that were used in practice rating were identical in each of the training conditions. Following this, the trainees received feedback on their ratings. The specific nature of the information the participants received with feedback was dependent upon the training condition. A question-answering phase addressed any remaining uncertainties the trainees may have concerning the subject matter of their respective experimental condition. Finally, a summary of each training perspective was provided.

Dependent Measures

A 7-item, 7-dimension interviewer rating questionnaire from Hauenstein (1987) was utilized at the pretraining and postraining measurement occasions, as well as within the three experimental conditions (see Appendix B). In addition,

this same form was used in the pilot study to develop the target scores. The 7 performance dimensions on the rating form, which are similar to Borman (1977), included: rapport building, organization of the interview, questioning skill, relevance of questions, company and job preview, answering the applicant, and an overall evaluation. Consistent with past performance appraisal research, the overall evaluation dimension was not included in the calculation of accuracy scores. This was because the overall evaluation dimension does not include intentionally independent job performance information. The rating format consisted of a 7-point Likert scale. While the items themselves are behavioral in nature, the 7, 4, and 1 anchors on the rating scale represented the three main categories of performance effectiveness (effective, average, ineffective).

The four major deviational components of accuracy described by Cronbach (1955) were used as dependent variables in the training study. These original scores have been utilized frequently in performance appraisal research over the past several years (e.g. Becker & Cardy, 1986; Murphy & Balzer, 1986; Murphy, Balzer, Kellam, & Armstrong, 1984; Pulakos, 1986). Cronbach (1955) argued that previously-used distance-squared indices included a diversity of useful information concerning accuracy in interpersonal perception. He proposed that the distance between ratings and target scores should be decomposed into four separate component accuracy scores. Each component, similar to an analysis of variance approach, was designed to represent a different portion of the distance between behavior ratings and target scores. Elevation expresses the differential grand mean variance, differential elevation represents the differential main effect of ratees, stereotype accuracy expresses the differential main effect of rating dimensions, and differential accuracy represents the differential ratee-dimension interaction. Accuracy and the value of these scores are inversely

proportional. In addition, the Cronbach, et al. (1972) generalizability theory was used to calculate the indices of rater reliability.

Procedure

Thirty training sessions consisting of from 2 to 6 trainees were conducted in a classroom setting. The average number of trainees in each of the 3 types of trainings was approximately 4. An effort was made to balance the number of idiosyncratic and normative raters, and the number of males and females, in each training session. Each of the three types of training procedures lasted for 90 minutes. Each of the seven training components was approximately the same duration of time across each type of training.

Three trainers were used in the study (1 male, 2 females). The trainers were blind as to the hypotheses of the study and the rating aptitude of trainees. To help control for experimenter effects, 2 trainers were assigned to a particular training type and alternated in leading the sessions for that training type. Therefore, each of the 3 trainers led 2 types of training sessions. The trainers participated in a one week-long series of trainings and practice sessions which closely followed the training protocols included in Appendices D through F. To help control for any potential gender biases, 1 male and 1 female trainer conducted the frame-of-reference experimental sessions. In turn, that male and the other female trainer ran the rater error experimental sessions. The two females conducted the control trainings. The male trainer ran 11 total sessions, one female ran 10 sessions, and the other female ran 9 sessions. Across all sessions, each of the trainers closely followed the training scripts or protocols included in Appendices D through F. An analysis of covariance (ANCOVA) was used to assess the differential impact of the trainers on the measures of accuracy. No trainer effects were found.

Across the three types of trainings, one randomly-ordered triad of interview segments (effective, average, and ineffective performance) was rated before training, a different triad was used within each training condition for the practice rating, and another was rated after the intervention phase was completed. In other words, all training groups viewed the same randomly-ordered triad at the same points in time described above. Across each of the three triads, an effort was made to balance a broad range of interviewer effectiveness.

In addition to the pretest and posttest accuracy measures, three other measures were used to gather information that might be relevant to the study's hypotheses. A job knowledge test, a rating aptitude retest, and two training tests were given to the subjects. All three measures were given to trainees in all three types of groups at the same stages of the training session (see Appendix C).

First, an interviewing knowledge test was constructed with two scales to help assess the effects of rater knowledge on rater accuracy. The first scale consisted of 15 multiple-choice test items which focused on the general aspects of the interview process from the interviewer and the applicant perspective. The second scale consisted of 10 self-report Likert items designed to assess the differential interviewing experiences of the subjects. Both interviewing knowledge scales were administered before the pretest. The 10-item self-report scale (Cronbach's alpha=.78) was much more reliable than the 15-item multiple choice items (Cronbach's alpha=.17). Therefore, the 10-item interviewing scale was used in the subsequent analyses.

Second, the same 100-item PRF rating aptitude test that was used to identify normative and idiosyncratic raters was administered after the posttest

rating. This aptitude retest was used to investigate any systematic changes in rating style as a result of the trainings.

Third, all trainees responded to a 10-item frame-of-reference training test and a 10-item rater error training test after the PRF retest. These training content tests were used to help verify that trainees were in fact acquiring pertinent information related to the training perspectives under study. The interviewing test, the PRF test, and the two training tests are included in Appendix G.

Each pretraining phase lasted approximately 25 minutes. Each posttraining phase lasted approximately 45 minutes. Trainees had a 10-minute break following the practice rating component, and one 10-minute break before the posttest performance rating. Therefore, the total time for the rater training study session, including both the training and the measurement phases, was 3 hours.

Results

Elevation, differential elevation, stereotype accuracy, and differential accuracy scores were calculated from the pretest and posttest performance ratings from each of the three training conditions. In addition, each subject's results for the interviewing test and the two training tests were computed. The means and standard deviations of these scores for each rating aptitude and training are presented in Tables 1 through 3. Lower values of the accuracy scores indicate higher levels of accuracy.

Identifying Normative and Idiosyncratic Raters at Pretest

Accuracy Before the analyses on the effects of the training procedures were conducted, pretest accuracy measures across rating aptitude were examined. The assumption that normative raters would be more accurate than idiosyncratic raters at pretest is part of the perspective that guides hypotheses 2a and 3a. This assumption was assessed using a one-way (rating aptitude) analysis of variance (ANOVA) on the pretest ratings of all three training conditions. Results of this ANOVA were only significant for differential accuracy ($F(1,127)=7.83$, $p<.01$). The mean difference in differential accuracy that was discovered was not in the expected direction. Idiosyncratic raters were more accurate at pretest ($M=.684$) than normative raters ($M=.795$). The other 3 accuracy scores showed no systematic mean differences at pretest between the two rating aptitude groups.

In terms of accuracy, the whole sample of raters appeared relatively homogeneous at pretest. Therefore, the rating aptitude-rater training interaction effect on rating accuracy included in hypotheses 2a and 3a may not be found because subjects were not differentiated on rating aptitude at pretest.

Reliability The assumption underlying the reliability hypotheses in this study (hypotheses 2b and 3b) was that normative subjects would be more reliable than the idiosyncratic participants at pretest. Intraclass rater reliability coefficients were calculated from the Cronbach, et al. (1972) generalizability formulae described in Cardinet, et al. (1976). Utilizing an ANOVA framework, interrater reliability can be indicated by the proportion of variance that is attributed to raters' systematic use of rating dimensions across ratees (Mitchell, 1979). When rating dimensions are used as the facet of differentiation and raters as the facet of generalization, the resulting intraclass correlation estimates the generalizability of the dimension ratings provided by a rater in a particular study's setting (cf. Hauenstein & Alexander, in press). To calculate the reliability estimates of normative and idiosyncratic raters at pretest, estimates of variance associated with each facet were obtained by means of two separate (one for the normative group, one for the idiosyncratic group) 6 (rating dimensions) by 3 (performance vignettes) ANOVAs.

When estimating interrater reliability via generalizability analysis, the effect of the number of raters is analogous to the effect that a number of items has on the reliability of a test. In classical test theory, a large number of items increases estimates of reliability. In generalizability theory, a large set of raters increases estimates of reliability. However, interrater reliability is conceptualized most commonly as the reliability between only two raters (Saal, Downey, & Lahey, 1980). Therefore, generalizability theory estimates of reliability using a large set of raters usually result in values which do not differentiate reliability levels effectively. In essence, a ceiling effect masks differences when there are a large number of raters. Fortunately, generalizability theory allows for generalization over any chosen number of

levels in the facet of generalization (raters). Thus, adjusted intraclass correlations generalized over two raters were used in this study to interpret the reliability differences between groups (Hauenstein & Alexander, in press). The unadjusted and adjusted pretest and posttest reliability coefficients for each rating aptitude and training condition are presented in Table 4. Unfortunately, ANOVAs cannot be conducted on these mean levels of interrater reliability.

The reliability coefficients of both rating aptitude groups were not different (M=.68). The rating aptitude-rater training interaction effect on reliability included in hypotheses 2b and 3b may not be found because the pretest reliability of normative raters did not differ significantly from idiosyncratic raters.

Training/Aptitude Effects on Accuracy

The effects of each of the training procedures on each accuracy component was investigated via separate repeated 2 by 3 ANOVAs. The two levels of aptitude (normative and idiosyncratic) and three levels of training (frame of reference, rater error, and control) were crossed, with the repeated factor being the measurement occasion (pretest and posttest). Tukey's HSD post hoc comparisons test was used when significant effects not included in the hypotheses were found.

Elevation Accuracy Hypothesis 1 predicted that in terms of elevation, normative and idiosyncratic rater error training participants would increase in accuracy from pretest to posttest relative to the other training groups. However, the expected measurement occasion-training interaction only approached significance ($F(2,123)=2.80$, $p<.10$). Summary information for the elevation ANOVA is presented in Table 5. Elevation scores of each training condition were examined to determine whether rater error training subjects increased in accuracy from pretest to posttest. The one-tailed, paired t-test revealed that only

frame-of-reference training subjects significantly changed their elevation accuracy scores from pretest to posttest ($t(40)=2.20$, $p<.05$). Across rating aptitude groups, frame-of-reference training subjects decreased in accuracy from pretest ($M=.391$) to posttest ($M=.569$). The other across rating aptitude training means, which were not significant, were in the expected direction. Rater error training subjects increased in accuracy from pretest ($M=.525$) to posttest ($M=.416$). Control training subjects decreased in accuracy from pretest ($M=.554$) to posttest ($M=.627$).

The rater training's effect on elevation accuracy information is displayed graphically in Figure 1. Elevation scores were examined at posttest to determine if rater error training subjects were more accurate than frame-of-reference or control training subjects. T-tests were used to test the a priori hypothesis that rater error training subjects would be more accurate than frame-of-reference training subjects. The one-tailed t-test revealed a significant posttest difference between the rater error training and frame-of-reference training groups ($t(84)=1.90$, $p<.05$). Rater error training subjects were more accurate ($M=.414$) than the frame of reference subjects ($M=.570$) at the posttest measure of elevation accuracy. The one-tailed t-test also revealed a significant difference between the rater error training and control training groups ($t(86)=-2.42$, $p<.01$). Rater error training subjects were more accurate ($M=.414$) than the control training subjects ($M=.626$) at the posttest measure of elevation accuracy. Therefore, rater error training subjects were most accurate in terms of elevational accuracy at posttest.

A one-way (training) anaylsis of covariance (ANCOVA) with pretest scores as the covariate was used to determine whether pretest training differences in elevation accuracy masked a rater error training effect on accuracy in the

original ANOVA. Controlling for pretest differences in elevation accuracy, a significant training effect was revealed ($F(2,125)=3.15$, $p<.05$). The contrast between rater error training and the other two trainings also was significant ($F(1,125)=5.80$, $p<.05$).

Thus, hypothesis 1 was partially supported. Rater error training resulted in higher levels of elevation accuracy in comparison to frame-of-reference and control training. However, evidence that rater error training subjects specifically increased in accuracy from pretest to posttest was only suggestive. Frame-of-reference training subjects were found to have decreased in elevation accuracy.

Stereotype Accuracy Hypotheses 2a and 3a stated that the rating aptitude-training interaction would occur with stereotype accuracy. However, the expected rating aptitude-training interaction from pretest to posttest was not significant ($F(1,123)=0.30$, $p>.10$). Summary information for the stereotype accuracy repeated measures ANOVA is presented in Table 6. The rating aptitude-training interaction for stereotype accuracy included in hypotheses 2a and 3a was not supported.

However, the measurement occasion-training interaction ($F(2,123)=7.84$, $p<.01$) and the measurement occasion-rating aptitude interaction ($F(1,123)=4.25$, $p<.05$) were significant. Post hoc analyses discovered a significant difference between frame of reference training and each of the other training types (Figure 2). Frame-of-reference training subjects ($M=.460$) were more accurate than rater error training subjects ($M=.711$) and control training subjects ($M=.634$).

Post hoc analyses also indicated a significant difference between the normative and idiosyncratic rating aptitude groups. Idiosyncratic participants were more accurate ($M=.553$) than normative participants ($M=.661$) at posttest (Figure 3).

Differential Accuracy The predicted three-way measurement occasion-rating aptitude-training interaction with differential accuracy was not significant ($F(2,123)=2.28$, $p>.10$). Therefore, hypotheses 2a and 3a were not supported. These results, in conjunction with the non-significant aptitude-training interaction finding with stereotype accuracy, clearly do not support hypotheses 2a and 3a. Summary information for the differential accuracy ANOVA is presented in Table 7.

Similar to stereotype accuracy, the measurement occasion-training interaction ($F(2,123)=8.31$, $p<.01$) and the measurement occasion-rating aptitude interaction ($F(1,123)=5.18$, $p<.05$) were significant. Post hoc analyses indicated that both frame-of-reference training and the control training differed significantly from rater error training at the posttest measure of differential accuracy (Figure 4). Frame-of-reference training subjects ($M=.497$) were more accurate than the rater error training subjects ($M=.711$). Similarly, control training subjects ($M=.556$) were more accurate than the rater error training subjects ($M=.711$).

Post hoc analyses indicated that the pretest difference in normative and idiosyncratic raters was non-significant at posttest (Figure 5). Idiosyncratic raters were not more accurate than normative raters at the posttest measure of stereotype accuracy.

Training/ Aptitude Effects on Reliability

Rating aptitude-training interactions were hypothesized for interrater reliability in hypotheses 2b and 3b. The interrater reliability estimates of normative and idiosyncratic raters in each of the training conditions were calculated by obtaining estimates of variance via 12 separate (one ANOVA for each

pretest and posttest experimental cell) 6 (rating dimensions) by 3 (ratees) ANOVAs. Table 4 presents this information.

Normative and idiosyncratic rater error training subjects displayed the hypothesized aptitude-training interaction, although normative and idiosyncratic subjects did not differ at pretest as hypothesized (Figure 6). Unfortunately, the idiosyncratic raters had a higher level of interrater reliability than the normative raters at pretest. The reliability of the normative rater error training participants decreased from .76 to .59. The reliability of the idiosyncratic rater error training subjects remained relatively unchanged (reliability decreased slightly from .77 to .72). Hypothesis 3b was only partially supported.

Both normative and idiosyncratic frame of reference subjects increased in interrater reliability at about the same magnitude from pretest to posttest (Figure 7). The reliability of normative frame of reference subjects increased from .66 to .85. The reliability of idiosyncratic frame of reference subjects increased from .54 to .72. Therefore, hypothesis 2b was not supported.

The reliability of the normative control training participants increased from .63 to .72. The reliability of the idiosyncratic control training participants decreased from .75 to .63. Reliability information for the control training group is displayed graphically in Figure 8.

Addenda

No hypotheses were put forth concerning differential elevation. In examining the ANOVA results (Table 8), no between subjects effects were significant. The within subject univariate test of measurement occasion-training interaction was significant ($F(2,123)=3.24$, $p<.05$). Post hoc analyses revealed a significant difference between the frame-of-reference and control training

groups. Frame of reference subjects (M=.482) were more accurate than control training subjects (M=.626) at the posttest measure of differential elevation. These training effects are displayed graphically in Figure 9.

Supplementary information was also collected to aid in the verification and explanation of this study's hypotheses. One assumption in the operationalization of rating aptitude was that normative raters would possess more job knowledge than idiosyncratic raters. Similarly, Hauenstein and Alexander (in press), and Hauenstein and Lord (1989), found positive relationships between rating aptitude and measures of IQ and performance on cognitive tasks. However, the results of the interviewing knowledge test did not differ significantly between the rating aptitude groups ($F(1,123)=0.10$, $p>.10$). Normative raters' responses had a mean of 15.69, while idiosyncratic raters' responses possessed a mean of 15.40. Cronbach's alpha for the 10-item interviewing scale used was .78. As would be expected from this result, rater knowledge was not a significant effect in any of the ANOVAs or ANCOVAs conducted with the 4 accuracy measures. This finding further suggests that raters across the 2 rating aptitude groups did not differ in rating aptitude.

Two 10-item tests were constructed for this study to help assess the raters' acquisition of information in frame-of-reference training and rater error training. However, very little can be learned from the results of the frame-of-reference training test. The mean number of correct responses across all 6 of the experimental cells were on a small range from 6.23 to 7.05. Cronbach's alpha was .21.

Concerning the rater error training test, rater error training subjects scored a mean of 8.31 correct responses, while subjects in the other two trainings had a combined mean score of 4.09. The rater error training

participants scored significantly higher than frame-of-reference training subjects ($t(84)=-15.24$, $p<.01$) and control training subjects ($t(86)=12.58$, $p<.01$). Cronbach's alpha for this test was .70. Therefore, the results concerning hypotheses 3a and 3b may be attributed to the rater error training subjects' acquisition of training information which was not acquired by participants in the other two groups. Not suprisingly, the scores of normative rater error training subjects did not differ significantly from the scores of idiosyncratic rater error training subjects ($t(43)=-1.14$, $p>.10$).

The 100-item PRF test used to identify normative and idiosyncratic raters was readministered in the posttraining phase. This retest provided an opportunity to assess the reliability of the rater aptitude test and to more specifically examine this study's rating aptitude results. Test-retest reliability may be approximated by the PRF results of the control training subjects. Less than half of the normative and idiosyncratic control subjects (48% and 46%, respectively) maintained their original rating aptitude classification which was measured approximately two weeks prior to the retest (Table 9). These results suggest that the rating aptitude test operationalization in this study possessed a low level of reliability.

Contrary to what was implied by this study's hypotheses, a larger amount of normative rater error trainees (82%) maintained their normative rating style than normative frame of reference trainees (60%). In addition, 35% of idiosyncratic rater error trainees improved their rating style to the normative classification, while only 14% of idiosyncratic frame of reference trainees improved to normative classification. As represented in Table 9, these results may be due to the rater error training's effect on threshold. When comparing rater error training results with the other two trainings, the normative rater

error training subjects showed the smallest decrease (18%) in threshold (versus 35% in frame-of-reference training and 47% in the control training). When comparing rater error training results with the other two trainings, the idiosyncratic rater error training subjects displayed the largest increase (65%) in threshold (versus 33% in frame-of reference training and 36% in the control training). Rater error training seemed to have a similar positive effect on both overall deviational accuracy (elevation accuracy) and overall deviational rating aptitude (threshold). However, rating aptitude changes must be interpreted cautiously because alterations in rating aptitude may be due to a regression to the mean effect (Cook & Campbell, 1979).

This study involved the use of 3 different trainers, 2 within each of the 3 training conditions. The differential impact of a trainer was investigated using an ANCOVA with the covariate trainer factor on each of the 4 accuracy components. The main effect of trainer was not significant at the .05 level.

Discussion

Overall, the results of this study did not support a rating aptitude-rater training interaction. In terms of frame-of-reference training, idiosyncratic raters did not become significantly more accurate or reliable while normative raters simultaneously maintained higher pretest levels of accuracy and reliability. Therefore, hypotheses 2a and 2b were not supported. On the whole, however, frame-of-reference training did consistently lead to the most accurate and reliable performance ratings. Frame-of-reference trainees only decreased in the elevation component of accuracy.

As a result of rater error training, normative raters did not become less accurate while idiosyncratic raters simultaneously maintained lower pretest levels of accuracy. Therefore, hypothesis 3a was not supported. Overall, rater error training did lead to the least accurate performance ratings, except on the measure of elevation. The normative rater error training group did decrease in their level of interrater reliability, while the idiosyncratic rater group simultaneously maintained their general pretest level of reliability. Unfortunately, the pretest level of interrater reliability for the idiosyncratic group was not lower than the pretest level of interrater reliability of the normative group. Therefore, there was only partial support for hypothesis 3b.

Similarly, there was partial support for hypothesis 1. Raters who participated in rater error training did possess the highest levels of posttest elevation accuracy. However, this difference was more attributable to an increase in elevation accuracy in the other training groups, rather than to rater error trainees' increase in accuracy from pretest to posttest. This finding suggests that the overall mean rating style of a rater is affected by rater error training in a different manner than by frame-of-reference or control training.

Results of the retest of rating aptitude also were convergent with this finding. When comparing rater error training results with the other two trainings, the idiosyncratic rater error training subjects displayed the largest increase in the threshold accuracy component of rating aptitude. Rater error training seemed to have a similar positive effect on both overall deviational accuracy as measured by elevation accuracy and overall deviational rating aptitude as measured by threshold. The effect on the level of elevation accuracy was expected since rater error training participants are taught to focus on their own rating behavior in reducing the incidence of rating errors. In contrast to frame-of-reference training, rater error training does not focus as specifically on dimension and ratee information which are represented by the other three components of accuracy.

By definition, idiosyncratic raters differ substantially from the norm in regards to their implicit theories concerning the job to be rated. This difference from the norm commonly results in a rating style that is less accurate and reliable (Hauenstein & Alexander, in press). Contrasting with a priori expectations, normative raters were not consistently more accurate and reliable than idiosyncratic raters at pretest. This result all but eliminates interpretations of the effects of rating aptitude in this study.

Several findings may be used as evidence that the whole group of raters were relatively homogeneous. First, the two rating aptitude groups did not differ on three of the four components of accuracy at pretest. Second, the overall levels of pretest interrater reliability for the normative and idiosyncratic raters were identical. In addition, other tests which are related to knowledge and performance on cognitive tasks should have also differentiated normative subjects from idiosyncratic subjects (Hauenstein & Alexander, in

press). However, the rating aptitude groups did not differ on the measure of job knowledge. Also, the results of the rater error and frame-of-reference training tests did not differ between normative and idiosyncratic raters.

There are five possible reasons for this unexpected lack of differentiation between rating aptitude types: 1) the cognitive abilities of the raters were homogeneous, 2) the raters' implicit theories about the interviewer were homogeneous, 3) the novelty of the rating task minimized the effect of rating aptitude differences, 4) the robustness of the trainings minimized the effect of rating aptitude differences, and 5) the measurement and identification of rating aptitude was unreliable.

First, the general cognitive ability of this study's sample may not have differed substantially across raters. In this way, this sample may have differed from the other studies that found significant effects for rating aptitude. While Hauenstein and Alexander (in press), and Hauenstein and Lord (1989), conducted their rating aptitude studies at universities with open admissions policies, this study took place at a university which utilizes a specific set of admissions standards. Therefore, this study's participants may have possessed a smaller range of cognitive ability, and thus a smaller range of rating aptitude. Such a small range of rating aptitude would be problematic in the identification of idiosyncratic raters who differ substantially from normative raters. It is not suprising that a true idiosyncratic group was not found in this study if the sample of raters were homogeneous. Results from Hauenstein and Foti (1989) suggested that it may be difficult to differentiate rating aptitudes in a homogeneous group of raters who possess approximately equal amounts of knowledge relative to the job. For example, only 2 out of 23 raters in a police supervisor group were identified as idiosyncratic raters in the Hauenstein and Foti study.

A similar factor which may have contributed to the lack of normative and idiosyncratic differences concerns the use of interviewer as the job to be evaluated. The job of interviewer may have been more unfamiliar to the group of subjects than was expected. This may have hindered any manifestation of rating aptitude effects at pretest. The majority of recent performance appraisal studies involve the evaluation of a lecturer by college student subjects (e.g. Murphy & Balzer, 1986; Steiner & Rain, 1989). The performance appraisal studies investigating rating aptitude have also involved college students evaluating the performance of lecturers (Hauenstein & Alexander, in press). The college student subjects in these studies probably possessed a large range of differentiated implicit theories regarding the performance of a lecturer. This large range of implicit theories leads to a large range of rating aptitude amongst the subjects. Therefore, normative and idiosyncratic raters may be more readily identified and the effects of rating aptitude may be observed more easily. In contrast, the college student subjects in this study may have possessed a relatively small range of undifferentiated implicit theories regarding the performance of an interviewer. Therefore, their interviewer rating aptitude may have been undifferentiated. The lack of normative and idiosyncratic differences on the interviewing knowledge test supports this position.

The novelty of the videotape rating task is the third factor that may have contributed to the masking of rating aptitude differences. Unfamiliarity of the rating task may have accounted for more of the variance in ratings at pretest than rating aptitude. The experimental subjects may have had to attend and adjust to the rating task before the relevant true systematic variance of rating aptitude could affect their ratings. For example, a few subjects communicated to the experimenters that they initially found it difficult to focus on the

interviewer because of the dyadic context of the interview presented on the videotapes. Therefore, while the subjects may not have differed in the desired direction at pretest, they may have been somewhat differentially affected by some of the the training content. In addition, the robustness of the training procedures cannot be discounted. The strength of the training procedures may have overshadowed any potential effects of rating aptitude. If this were the case, research concerning the relation between rating aptitude and rater training may be unneeded. Unfortunately, at this point in time the results of this study cannot be compared with other studies to investigate this possibility. Virtually no other studies have tested the relation between individual rating aptitude differences and the effects of performance appraisal training.

Finally, it may simply be that the method of measurement of rating aptitude was unreliable. Test-retest reliability may be approximated by the PRF results of the control training subjects. Although a null training group would have provided a truer estimate of test-retest reliability, the control training was designed to have a minimal impact on the rating aptitude raters. Less than half of the normative and idiosyncratic control subjects maintained their original rating aptitude classification. Although relevant systematic variance due to the control training may have contributed in part to the change in rating aptitude, such a large change in rating aptitude was suprising. A history threat to internal validity (Cook & Campbell, 1979) may also explain the change in rating aptitude, but it is doubtful that many of this study's subjects acquired significant information which affected their rating aptitude of interviewers between test and retest. Therefore, the control training subjects' change in rating aptitude may consist of random error. This would lead to the conclusion that the identification of normative and idiosyncratic subjects was unreliable in

this study. This contention of unreliability does contrast with the previous work utilizing the PRF test to operationalize rating aptitude (e.g. Hauenstein & Alexander, in press). The previous research did find reliable rating aptitude effects on measures of rating accuracy, interrater reliability, and cognitive ability.

In summary, it is plausible that all five of these factors may have contributed to the lack of differences between normative and idiosyncratic raters. Raters may have been more similar than expected in terms of general cognitive ability and specific job knowledge. Thus, their implicit theories that directly relate to rating aptitude may have been similar. Also, any potential rating aptitude differences that did exist may have been eliminated by the novelty of the rating task and the robustness of the procedure. Since this is the first study that specifically examined the relation between rating aptitude and rater training, it is difficult to compare the training robustness proposition to other experiments. In the end, these four factors may have interacted to produce unreliability in the measurement of rating aptitude.

The strong training effects found with frame-of-reference training in this study did effectively replicate results from past frame of reference studies (e.g. Pulakos, 1984; 1986). In terms of accuracy in evaluating ratees and using dimensions (the differential elevation and stereotype accuracy scores, respectively), frame-of-reference training proved more effective than both the control and rater error trainings. Because training format or method of presentation was balanced across trainings, it can be concluded specifically that the frame-of-reference approach results in the highest amount of accuracy in evaluating ratees and using rating dimensions. Frame-of-reference training improves differential elevation and stereotype accuracy above and beyond the

control training format of general examples and practice/feedback. Thus, the true advantage in frame-of-reference training is probably the specificity of the examples and practice/feedback which effectively convey performance standards. This is similar to Athey and McIntyre's (1987) conclusion that the provision of rating standards and behavioral examples were responsible for improvements in rating accuracy. By utilizing true or target scores in examples and feedback that reflect some level of veridicality, a rater more readily learns through direct experience how to evaluate different ratees and use different rating dimensions. In essence, the frame-of-reference training session is a microcosm that includes an efficient model of the normal standard acquisition process.

Interestingly, the results of frame-of-reference training did not differ from control training at the level of differential accuracy. This finding suggests that the training components that were similar in both trainings resulted in increasing differential accuracy. Thus, it may be concluded that just examples and practice in observing and rating behavior is beneficial to the integration of dimension and ratee information. That is, multiple experiences with job performance will help improve accuracy as measured by differential accuracy. It seems that frame-of-reference training, which includes examples and practice specifically tied to true or target scores, positively affects the accuracy of dimension and ratee judgements separately. In contrast, examples and practice, without the use of true or target scores, positively affects the accuracy of the rater's integration or combination of dimension and ratee information.

Two rating aptitude results of statistical significance were found in this study. Interpretation of their meaning is difficult, given that it has been concluded above that little true difference exists between the normative and

idiosyncratic raters. That is, the variance in performance ratings attributed to rating aptitude in fact might be bias or random error. Post hoc analyses indicated that normative raters were less accurate than idiosyncratic raters at posttest. It was also indicated that at the pretest measure of differential accuracy, idiosyncratic raters were more accurate than normative raters. These rating aptitude effects contrast with the work of Hauenstein and Alexander (in press), and Hauenstein and Lord (1989).

While this study cannot extend the pertinent rater training research literature by commenting on differential aptitude effects, some implications can be drawn from what is known about this study's rater sample. The whole group of raters can be characterized as possessing a limited range of rating aptitude. Further, this group did not have a large amount of experience in evaluating the particular job of concern. This is in contrast to, for example, the group of experienced patrol officer supervisors in Hauenstein and Foti (1989). This study's rater sample is probably similar to a set of potential managers who are new to an organization, and who are just beginning their careers in a particular area of expertise. Because of selection standards, these employees are somewhat homogeneous in terms of rating aptitude. Since the managers are novices in a particular area, they also do not have a large amount of experience in evaluating the specific job of concern. It may be concluded that frame-of-reference training could be particularly effective for first-time managers within an organization because the characteristics of these managers may be similar to this study's rater sample.

Another issue in regards to the implications for this study is the actual degree to which these findings may be applied and transferred to the workplace. The criteria for the effectiveness of the trainings was accuracy that was

computed from the subjects' ratings of videotapes. However, the observation of performance in the work setting may require more effort and be complicated by multiple competing tasks. In the workplace, the effects of frame-of-reference training might be diffused because a manager would not be able to focus on specific aspects of performance and the use performance standards to evaluate behavior.

In addition, true or target scores for performance examples and practice/feedback would probably not be available in the applied setting. This situation may be problematic because target scores which reflect reality are significant to the improvement of differential elevation and stereotype accuracy. Some form of organizationally-defined standards will have to replace empirically-defined standards when conducting frame-of-reference training in the applied setting. Research should continue on the development of performance appraisal standards within organizations. This study's findings also found that practice in observing and rating performance improves measures of differential accuracy. More frequent frame-of-reference trainings and more frequent rating practice sessions represent two techniques which can improve performance appraisal accuracy.

References

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Level-of-processing theory and social facilitation theory perspectives. Journal of Applied Psychology, 72, 239-244.

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

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

Bernardin, H. J. (1979). The predictability of discrepancy measures of role constructs. Personnel Psychology, 32, 139-153.

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

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

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

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

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

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

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

Borman, W. C. (1983). Implications of personality theory and research for the rating of work performance in organizations. In F. Landy, S. Zedeck, and J. Cleveland (Eds.), Performance measurement and theory. Hillsdale, New Jersey: Erlbaum.

Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an army officer sample. Organizational Behavior and Human Decision Processes, 40, 307-322.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 13, 119-135.

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

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

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

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

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

Hauenstein, N. M. A. (1987). A process approach to ratings: The effects of ability and level of processing on encoding, retrieval, and rating outcomes. Unpublished doctoral dissertation, University of Akron, Akron, OH.

Hauenstein, N. M. A., & Alexander, R. A. (in press). Rating ability in performance judgments: The joint influence of implicit theories and intelligence. Organizational Behavior and Human Decision Processes.

Hauenstein, N. M. A., & Foti, R. J. (1989). From laboratory to practice: Neglected issues in implementing frame-of-reference rater training. Personnel Psychology, 42, 359-378.

Hauenstein, N. M. A., & Lord, R. G. (1989, April). The effects of individual differences on the encoding of performance information. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Boston, Massachusetts.

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

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In B. Shaw & L. L. Cummings (Eds.), Research in organizational behavior, (vol. 5). Greenwich, CT: JAI Press.

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

Jackson, D. N. (1972). A model for inferential accuracy. The Canadian Psychologist, 13, 185-194.

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

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

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

London, M., & Hakel, M. D. (1974). Effects of applicant stereotypes, order, and information on interview impressions. Journal of Applied Psychology, 59, 157-162.

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

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 86, 376-390.

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

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

Nathan, B. R., & Alexander, R. A. (1985). The role of inferential accuracy in performance rating. Academy of Management Review, 10, 109-115.

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

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

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

Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. Journal of Applied Psychology, 74, 136-142.

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

Thorndike, E. L. (1920). A constant error in psychological ratings. Journal of Applied Psychology, 4, 25-29.

Appendix A

Interviewer Behaviors Represented on the Stimulus Videotapes as Categorized by Hauenstein (1987)

Effective Behaviors

1. Interviewer asked the applicant about specialized training.
2. Interviewer asked the applicant about relevant work experience.
3. Interviewer asked the applicant about colllege major.
4. Interviewer asked the applicant about college grade point average.
5. Interviewer provided a job preview to the applicant.
6. Interviewer showed the applicant's position location on an organizational chart.
7. Interviewer asked the applicant why he was interested in the new position.
8. Interviewer discussed certain aspects of the organization's fringe benefits.
9. Interviewer exchanged introductions with the applicant.
10. Interviewer arranged an interview with the executive to whom the applicant would directly report.
11. Interviewer asked the applicant why he was leaving his current position.
12. Interviewer stated how job performance was evaluated in the organization.

Average Behaviors

1. Interviewer asked the applicant about preferences for geographical location.
2. Interviewer asked the applicant about his limitations.
3. Interviewer asked the applicant about his most rewarding work experience.
4. Interviewer looked at applicant's resume during the interview.
5. Interviewer asked the applicant about his community involvements.
6. Interviewer set up an appointment between the applicant and the previous job holder.
7. Interviewer explained the purpose of the interview.

7. Interviewer explained the purpose of the interview.

8. Interviewer asked the applicant about his most difficult work experience.

9. Interviewer stated the organization's growth patterns.

10. At the end of the interview, the interviewer summarized the information the applicant had provided.

11. Interviewer asked the applicant about his college extracurricular activities.

12. Interviewer asked the applicant to repeat a response.

Ineffective Behaviors

1. Interviewer asked the applicant about his elementary school experiences.

2. Interviewer interrupted an applicant's response.

3. Interviewer offered the applicant a cigarette.

4. Interviewer asked the applicant about his reactions to former college professors.

5. Interviewer asked about the applicant's trip to the interview.

6. Interviewer forgot to arrange a meeting between the applicant and the supervisor.

7. Interviewer asked the applicant about books he read recently.

8. Interviewer refused to answer a question from the applicant.

9. Interviewer picked lint off his clothes during the interview.

10. Interviewer used the phone to call his secretary during the interview.

11. Interviewer asked the applicant about movies he had seen recently.

12. Interviewer left the room during the interview.

Appendix B

Below is a list of rating dimensions for evaluating the interviewer. Read each dimension. Then, rate the interviewer on that dimension by filling in an appropriate oval that most closely reflects your evaluation of the interviewer.

1. RAPPORT BUILDING: The extent to which the interviewer put the applicant at ease, built trust, and was socially desirable.

1	2	3	4	5	6	7
Poor			Average			Excellent

2. ORGANIZATION OF INTERVIEW: The extent to which the interview had a clear format which was followed.

1	2	3	4	5	6	7
Poor			Average			Excellent

3. QUESTIONING SKILL: The extent to which questions were tactfully presented.

1	2	3	4	5	6	7
Poor			Average			Excellent

4. RELEVANCE OF QUESTIONS: How useful the interviewer's questions were for assessing the knowledge, skills, abilities, and characteristics of the applicant.

1	2	3	4	5	6	7
Poor			Average			Excellent

5. JOB AND COMPANY PREVIEW: The extent to which knowledge concerning the company and the job for which the applicant has applied is appropriately conveyed.

1	2	3	4	5	6	7
Poor			Average			Excellent

6. ANSWERING THE APPLICANT: The demonstrated willingness for answering the applicant's questions and the extent to which the responses provided meaningful information.

1	2	3	4	5	6	7
Poor			Average			Excellent

7. OVERALL EVALUATION: Your general impression of the interviewer's overall performance in the interview.

1	2	3	4	5	6	7
Poor			Average			Excellent

Appendix C

Training Session Procedure Overview with Component Durations

FOR Training (3 hr.)	RE Training (3 hr.)	Control Training (3 hr.)
:10 Interviewing Test	:10 Interviewing Test	:10 Interviewing Test
:15 Pretest (segments 1,2,3)	:15 Pretest (segments 1,2,3)	:15 Pretest (segments 1,2,3)
:05 Perf. appraisal, Work Multidim., Behavior Obs. Lecture	:05 Perf. appraisal, Work Multidim., Behavior Obs. Lecture	:05 Perf. appraisal, Work Multidim., Behavior Obs. Lecture
:10 Gen. Job Behaviors and Dims. Disc.	:10 Gen. Job Behaviors and Dims. Disc.	:10 Gen. Job Behaviors and Dims. Disc.
:10 FOR Concepts Lecture and Disc.	:10 RET Concepts Lecture and Disc.	:10 Org. Psych. Lecture and Discussion
:20 Job Beh. Examples via FOR Concepts (6 dims x good, average, poor) (segments 4,5,6,7,8,9)	:20 Job Beh. Examples via RET Concepts (segments 4,5,6,7,8,9)	:20 Job Beh. Examples (segments 4,5,6,7,8,9)
:15 Practice Rating (segments 10,11,12)	:15 Practice Rating (segments 10,11,12)	:15 Practice Rating (segments 10,11,12)
:10 (Break)	:10 (Break)	:10 (Break)
:15 Rating Feedback via FOR Concepts	:15 Rating Feedback via RET Concepts	:15 Rating Feedback
:10 FOR Summary Discussion	:10 RET Summary Discussion	:10 Org. Psych. Discussion
:10 (Break)	:10 (Break)	:10 (Break)
:15 Posttest (segments 13,14,15)	:15 Posttest (segments 13,14,15)	:15 Posttest (segments 13,14,15)
:20 Aptitude Retest	:20 Aptitude Retest	:20 Aptitude Retest
:10 FOR, RET Tests	:10 FOR, RET Tests	:10 FOR, RET Tests

Appendix D

Frame-of-Reference Training Protocol

Materials
dress in semi-formal clothes (look professional)
training protocol for condition assigned to that session
overview of training conditions
VCR (monitor will be in rooms already)
connection cable for VCR to monitor
chalk
extra coding pencils
training packet (interviewing survey, rating forms, aptitude retest, training information survey)
orange opscans (enough for 4 opscans per intro. student)
rating justification sheets
blue data opscans
informed consent forms

using 1st half of training videotape-- larger VCR start:
smaller VCR start:

***Approximate the time durations on the training overview as you go along and conduct the session.

Introduction and Pretest

Distribute informed consent forms.

Thank you all for coming. This important study is investigating the ways in which we can train people to evaluate job performance. Today's experimental session will run about 3 hours in total. Because of the length of this session, I will be letting you all take breaks at certain points in the middle of the experiment. First, I'd like to hand out some of the materials will be using in this session.

Distribute training packet and blue opscan form.
Have subjects code their Soc Sec. number in lines 1-9,
code their condition on line 10 (FOR=1, RET=2, CON=3)
code responses to interviewing survey in columns A-Y in top box.

At this point, I would like you all to view 3 performance segments on the VCR. After each segment, I will give you time to make performance ratings. You will be evaluating the interviewer. He is seated behind the desk and is wearing glasses.

Show three segments, stop after each for rating.
Tell subjects to skip line 11.
Seg 1, lines 12-18; Seg 2, lines 19-25, Seg 3, lines 26-32.

Perf. Appraisal, Multidimensionality of Work, Beh. Observation Lecture

You all just completed 3 performance appraisals after watching 3 interview segments on the VCR. Some of you may have had experience in your past with this type of task. Managers and supervisors commonly evaluate the performance of their subordinates to insure that their group or organization is running the best way possible. Often, this information on the quality of an employee's performance is presented, in turn, to the employee. This feedback occurs so that the employee can become more aware of the ways in which they are performing well, and the ways in which that employee may not be performing well. This performance information that is presented to the employee (as you might imagine) is extremely important to the overall performance of the organization.

On the performance appraisal forms that you completed earlier, you might also remember that there were a number of performance dimensions which were to be used in rating the interviewer. You all, in essence, made an employee performance rating for each of these performance dimensions. That is, for each employee doing a particular job (in this case, the interviewer in each of the 3 segments) you made a handful of ratings. For each interview segment, you made 6 or 7 performance ratings. Well, why did you make all these ratings for each segment? Shouldn't an overall performance rating be enough? The answer is that one overall rating is probably not enough to capture the quality of performance of that employee. Most jobs are not unidimensional. An employee could be strong on certain rating dimensions, but poor on others. And as you might conclude, all this information is very important to the organization and to the feedback to the employee. Work performance is viewed largely as being multidimensional in nature. And this is why it is common to see quite a few rating dimensions on performance appraisal forms.

However, a good performance appraisal system cannot be built solely upon a rating form which sufficiently describes a job along a handful of rating dimensions. In order for performance ratings to give an accurate portrayal of an employee's performance, the rater must pay close attention to that employee and that employee's work performance. And because this work performance is variable and multidimensional, the person who is making the ratings needs to attend to as much of the employee's specific work behaviors as is possible. That is, in order for a supervisor to be a good rater of employee performance, he or she must be an attentive observer of what that employee does and says in the line of work. Being a good observer in making performance ratings for each job dimension is the first step towards accurate and useful performance evaluations.

General Job Behaviors and Dimensions Discussion

Now that I have talked about the general nature of the performance appraisal process, I'd now like to talk about the specific job which we have been concerned with, and will continue to focus upon. The job is an interviewer of job applicants at a mid-sized, technologically-oriented organization. Thinking about what you observed in those 3 interview segments, and using the knowledge and experience you possess about such interviewers of job applicants, what behaviors are commonly required in such a job?

Examples are below.

1. Communicate with selection people, to select potential interviewees, help in making selection decisions.

2. Contact interviewee and set up interview.
3. Gather knowledge about specific open position.
4. Set up meetings with past position holders, potential supervisors, potential coworkers.
5. Organize interview, write questions concerning education, training, experience.
6. Establish rapport with interviewee.
7. Give preview of the job, organization.
8. Answer interviewees' questions.

Now that we've identified a lot of the behaviors required of an interviewer of job applicants, let's go back to the original rating dimensions which you used. Most rating forms, like the ones you had in front of you, are put together after someone has taken into account behaviors required on the job. In constructing the rating dimensions, the job behaviors are basically grouped into major categories. The rating dimensions for the job of interviewer are:

Write general names on board.

1. Rapport building- The extent to which the interviewer put the applicant at ease, built trust, and was socially appropriate.
2. Organization of the interview- The extent to which the interview had a clear format which was followed.
3. Questioning skill- The extent to which questions were tactfully presented.
4. Relevance of questions- How useful the interviewer's questions were for assessing the knowledge, skills, abilities, and characteristics of the applicant.
5. Company and job preview- The extent to which knowledge concerning the company and the job is appropriately conveyed.
6. Answering the applicant- The demonstrated willingness for answering the applicant's questions and the extent towhich the responses provided meaningful information.
7. Overall evaluation- Your general impression of the interviewer's overall performance in the interview.

Do you have any questions about the definitions of each of the rating dimensions?

Training Concepts and Examples-- Lecture and Discussion

We have now discussed the behaviors generally required of the employee on the job, and we have discussed the rating dimensions that are used in evaluating that employee's performance. At this point, there are 2 ways in which we could go further in examining employee behaviors and rating dimensions, and increase the effectiveness of the performance evaluation.

One way is to further specify which employee behaviors should be rated on which specific rating dimensions. That is, match a potential employee behavior to its appropriate rating dimension. However, just knowing what behaviors are related to which job performance dimensions is probably not enough to promote accurate ratings. For example, it is not sufficiently helpful for a rater to know that when the interviewer answers a question, that behavior belongs in dimension number 6, Answering the Applicant.

Rater accuracy is dependent upon being more discerning about a behavior such as the interviewer answering the applicant. Rater accuracy can be fostered by the rater knowing what are good, average, and poor examples of job behavior within each job-relevant dimension. Then, a good rater can compare what they observe to these appropriate standards of behavior.

Any questions?

So what are some examples of good and poor behaviors in each of these dimensions?

Write examples on board.

I will now show you a good, average, and poor behavior example for each of the 6 main rating dimensions. With each example you see on the VCR, I will tell you the actual rating that behavior should have received if you were rating it, and I will also point out specific behaviors which caused that rating. During the course of these examples for each rating dimension, you may notice that some of the behaviors may be repeated and used with several dimensions. This should not be surprising, however, since as we talked about earlier, job performance is multidimensional in nature.

View videotape of good, average, poor examples for each of the 6 dimensions. Trainer should relate the target score for each example, followed by a behavioral rationale of that rating. For each dimension, ask group whether they saw the behavior, and ask whether they have questions about the behavior(s) which contributed to the rating.

Rapport Building
- good, 5, asked about cough, talked about weather, geographical preferences where applicant might live
- average, 4, summarized information at end of interview
- poor, 3, interviewer left office in middle of interview

Organization of the Interview
- good, 5, asked about training, then about relevant work experience
- average, 4, asked about extracurricular activities, and then how they would transfer to the job
- poor, 1, interviewer left office, long pause, asked about movies

Questioning Skill
- good, 6, asked about training, work experience, had follow-up questions
- average, 5, asked about geographical preferences of applicant, asked about limitations, questions did not go together
- poor, 4, asked applicant to repeat response

Relevance of Questions
- good, 6, asked job-related questions-- about training, work experience
- average, 4, asked about community involvements, how they might relate to the job
- poor, 1, asked about movies

Company and Job Preview
good, 6, stated how performance on the job was evaluated
average, 5, described the company's growth patterns
poor, 4, talked about weather, some talk about job location

Answering the Applicant
good, 6, answered questions about how performance was evaluated, how supervisors get performance information
average, 4, answered time question by setting up appointment with previous job holder
poor, 2, refused to answer applicant's question

As you can see, some behaviors and rating dimensions are easier to work with than others. By looking at this range of examples in this way, and then talking about them a bit, helps you to create an accurate frame of reference which you can use in making more accurate performance judgments.

Practice Rating

Using the information you've gathered about the performance appraisal process, I'm now going to show you 3 more interview segments so that you can become more acquainted with and practice some of the definitions and concepts we just talked about. Again, you will be rating the performance of the interviewer, the man with glasses seated behind the desk. In addition to coding your ratings on your opscan sheet, write out justifications for your evaluation of each of the 3 segments on the sheet provided. This will prove beneficial when we go back over your ratings together.

Have subjects write first name and Soc. Sec number on justification sheet.

Distribute numbered justification sheet for practice.

Show three segments, stopping after each for rating.
Seg 1, lines 33-39; seg 2, lines 40-46; seg 3, lines 47-53.

Feedback and Discussion of Practice Rating

When group is done, have them write their first name in pencil on top line of blue opscan.

Then collect opscans and give the group 10 min break.

Write ratings from opscans on board, ratee by dimension for each rater. Write down target scores for each of the 3 segments in the same format. Hand back opscans.

Target scores:

#1	4.1	3.5	3.7	3.2	3.8	4.5	3.7
#2	2.6	2.6	2.9	3.5	2.6	2.8	2.6
#3	5.6	4.8	4.7	3.9	4.1	5.0	4.8

These are your ratings of the 3 interview segments. And these are the true, or target performance ratings of the 3 interview segments you just saw. These ratings represent the true level of performance, the most accurate ratings for the performance you just viewed. These ratings are norms that we computed from a large sample of raters. You can see that there are some similarities, and some differences between your ratings and these true performance ratings. Our goal is to reduce the differences so that you all rate very similarly to the true performance ratings. Let's take a look at each of your ratings individually.

For each rater:

Ask rater to read or describe their reasons for the way in which they rated each of the 3 segments.
Ask group how on target that rater's behavioral rationale was for each of the 3 segments.
As the trainer, make sure each specific rating is discussed in its similarity or discrepancy from the true performance rating. Make sure reasons are identified for both the similarities and the differences.

Possible rationales mentioned by raters:
#1
asked why applicant was interested in the new position
discussed fringe benefits
talked about company's span of control
mentioned dental plan, after being reminded
#2
called secretary in middle of interview
indicated that he forgot to schedule a meeting
asked few questions of the applicant
asked short questions relating to projects the applicant was talking about
#3
introduced self
arranged interview with supervisor
explained approach to interviewing
talked to secretary before the interview
shook hands with the applicant

Any questions concerning the behaviors that led to the true performance ratings?

Any questions about the relation between your ratings and the true performance ratings?

Summary of Training

Soon, you will have one more opportunity to rate the performance of the interviewer. At this point, I would like to talk about the main points in our discussion on improving the quality of performance ratings.

First of all, when I mention frame of reference, or the learning of performance standards, what am I talking about? (Possessing job relevant behavioral examples of various levels of performance, within each job dimension)

What are some of the ways in which you learned more about the accurate frame of reference for the job of interviewer? (Talked about examples of behavior on the job, shown examples of interviewer behavior with true performance ratings and reasons for those ratings, practiced rating with these new performance standards, got feedback on how well you did in comparison to true performance ratings and reasons for those ratings)

What did you all get out of practicing? And then receiving feedback? (hopefully, you honed in on performance standards)

And the most important question, how can you improve the quality of your ratings? (observe all job relevant behavior, know the job-relevant rating dimensions, use the standards you've acquired for each job dimension, to rate job-related behaviors)

At this point, you all can take a 10 minute break. When you come back, you will give one last set of interviewer performance ratings, and then fill out a couple of surveys.

Posttest

Show three performance segments, stopping after each for rating.
Seg 1, lines 54-60; seg 2, lines 61-67; seg 3, lines 68-74.

Then read instructions for aptitude test.
Then tell subjects to complete aptitude retest, lines 75-174.

Then have subjects respond to the training information survey,
lines 175-194.

Sign extra credit materials, have intro. students fill out 4 orange opscans, with their name, coded ID#, and in seat number code 002.
Give them brief debriefing about purpose of the experiment and answer any questions.

Appendix E

Rater Error Training Protocol

Materials
dress in semi-formal clothes (look professional)
training protocol for condition assigned to that session
overview of training conditions
VCR (monitor will be in rooms already)
connection cable for VCR to monitor
chalk
extra coding pencils
training packet (interviewing survey, rating forms, aptitude retest, training information survey)
orange opscans (enough for 4 opscans per intro. student)
rating justification sheets
blue data opscans
informed consent forms

using the 2nd half of the training videotape-- larger VCR start:
smaller VCR start:

***Approximate the time durations on the training overview as you go along and conduct the session.

Introduction and Pretest

Distribute informed consent forms.

Thank you all for coming. This important study is investigating the ways in which we can train people to evaluate job performance. Today's experimental session will run about three hours in total. Because of the length of this session, I will be letting you all take breaks at certain points in the middle of the experiment. First, I'd like to hand out some of the materials will be using in this session.

Distribute training packet and blue opscan form.
Have subjects code their Soc Sec. number in lines 1-9,
code their condition on line 10 (FOR=1, RET=2, CON=3)
code responses to interviewing survey in columns A-Y in top box.

At this point, I would like you all to view 3 performance segments on the VCR. After each segment, I will give you time to make performance ratings. You will be evaluating the interviewer. He is seated behind the desk and is wearing glasses.

Show three segments, stop after each for rating.
Tell subjects to skip line 11.
Seg 1, lines 12-18; Seg 2, lines 19-25, Seg 3, lines 26-32.

Perf. Appraisal, Multidimensionality of Work, Beh. Observation Lecture

You all just completed 3 performance appraisals after watching 3 interview segments on the VCR. Some of you may have had experience in your past with this type of task. Managers and supervisors commonly evaluate the performance of their subordinates to insure that their group or organization is running the best way possible. Often, this information on the quality of an employee's performance is presented, in turn, to the employee. This feedback occurs so that the employee can become more aware of the ways in which they are performing well, and the ways in which that employee may not be performing well. This performance information that is presented to the employee (as you might imagine) is extremely important to the overall performance of the organization.

On the performance appraisal forms that you completed earlier, you might also remember that there were a number of performance dimensions which were to be used in rating the interviewer. You all, in essence, made an employee performance rating for each of these performance dimensions. That is, for each employee doing a particular job (in this case, the interviewer in each of the 3 segments) you made a handful of ratings. For each interview segment, you made 6 or 7 performance ratings. Well, why did you make all these ratings for each segment? Shouldn't an overall performance rating be enough? The answer is that one overall rating is probably not enough to capture the quality of performance of that employee. Most jobs are not unidimensional. An employee could be strong on certain rating dimensions, but poor on others. And as you might conclude, all this information is very important to the organization and to the feedback to the employee. Work performance is viewed largely as being multidimensional in nature. And this is why it is common to see quite a few rating dimensions on performance appraisal forms.

However, a good performance appraisal system cannot be built solely upon a rating form which sufficiently describes a job along a handful of rating dimensions. In order for performance ratings to give an accurate portrayal of an employee's performance, the rater must pay close attention to that employee and that employee's work performance. And because this work performance is variable and multidimensional, the person who is making the ratings needs to attend to as much of the employee's specific work behaviors as is possible. That is, in order for a supervisor to be a good rater of employee performance, he or she must be an attentive observer of what that employee does and says in the line of work. Being a good observer in making performance ratings for each job dimension is the first step towards accurate and useful performance evaluations.

General Job Behaviors and Dimensions Discussion

Now that I have talked about the general nature of the performance appraisal process, I'd now like to talk about the specific job which we have been concerned with, and will continue to focus upon. The job is an interviewer of job applicants at a mid-sized, technologically-oriented organization. Thinking about what you observed in those 3 interview segments, and using the knowledge and experience you possess about such interviewers of job applicants, what behaviors are commonly required in such a job?

Examples are below.

1. Communicate with selection people, to select potential interviewees, help in making selection decisions.

2. Contact interviewee and set up interview.
3. Gather knowledge about specific open position.
4. Set up meetings with past position holders, potential supervisors, potential coworkers.
5. Organize interview, write questions concerning education, training, experience.
6. Establish rapport with interviewee.
7. Give preview of the job, organization.
8. Answer interviewees' questions.

Now that we've identified a lot of the behaviors required of an interviewer of job applicants, let's go back to the original rating dimensions which you used. Most rating forms, like the ones you had in front of you, are put together after someone has taken into account behaviors required on the job. In constructing the rating dimensions, the job behaviors are basically grouped into major categories. The rating dimensions for the job of interviewer are:

Write general names on board.

1. Rapport building- The extent to which the interviewer put the applicant at ease, built trust, and was socially appropriate.
2. Organization of the interview- The extent to which the interview had a clear format which was followed.
3. Questioning skill- The extent to which questions were tactfully presented.
4. Relevance of questions- How useful the interviewer's questions were for assessing the knowledge, skills, abilities, and characteristics of the applicant.
5. Company and job preview- The extent to which knowledge concerning the company and the job is appropriately conveyed.
6. Answering the applicant- The demonstrated willingness for answering the applicant's questions and the extent towhich the responses provided meaningful information.
7. Overall evaluation- Your general impression of the interviewer's overall performance in the interview.

Do you have any questions about the definitions of each of the rating dimensions?

Training Concepts and Examples-- Lecture and Discussion

When making performance judgments based upon these rating dimensions, raters are often prone to making certain errors. It seems that those who use the rating forms and the rating dimensions just like you all did often let biases affect their ratings of an employee.

One bias is that raters are often guided by a central impression in making their ratings. The result is that a rater may rate an employee similarly across all of the rating dimensions. However, as we know, work performance is multidimensional. As was mentioned earlier, employees commonly may be performing well on certain dimensions, and not so well on other dimensions.

This central impression error is called halo error. It occurs when all the ratings for an employee across dimensions are unitarily high, or when all the dimension ratings are similarly low. The best way to combat halo error, or

central impression error, is to make sure that your dimension ratings for each performance segment include high ratings and low ratings.

The other bias that negatively affects performance ratings is that often raters are too lenient in their evaluations of employees. Managers and supervisors may often be too easy-going and not tough enough in their appraisals of their workers' performance.

This error is called leniency error. The best way to combat leniency error is by giving employees lower ratings.

Any questions?

So what are some examples of halo error in everyday life, when our central impression leads us to making overall judgments across a variety of dimensions or categories? (people who are close to us--friends, family, roommates; general impressions--first dates, actors/actresses, athletes, strangers on a street)

What are some examples of leniency error in everyday life? (people who are close to us--friends, family, roommates)

Why might a manager who is giving a performance rating take it easy on a worker? (friendship, don't want to hurt person's feelings, don't want person to be fired, feel that worker is reflection of the manager)

I will now show you 3 examples of halo error and 3 examples of leniency error. After each performance segment you see on the VCR, I will show you a set of ratings that represent the error.

View 3 videotapes of halo error (1, 7, 16). After each segment, show corresponding ratings exemplifying halo error. Trainer should relate the rationale for labelling each rating set as an example of an error, ask group whether they see the error in the ratings, and ask whether the group has any questions.

View 3 videotapes of leniency error (9, 12, 18). After each segment, show corresponding ratings exemplifying leniency error. Trainer should relate the rationale for labelling each rating set as an example of an error, ask group whether they see the error in the ratings, and ask whether the group has any questions.

#1	6	6	6	6	6	6	6
#7	4	4	4	4	5	4	4
#16	1	2	2	2	2	2	1
#9	7	6	7	6	6	7	7
#12	6	5	5	6	6	5	6
#18	6	6	7	5	6	6	7

As you can see, it takes a little work to get a feel for the definitions of these errors. Looking at these examples, and then talking about them a bit, should help you to grasp what these errors are. To improve your ratings of job performance, you need to avoid making these errors by spreading out your ratings for each performance segment, and by keeping your ratings lower on the rating scale.

Practice Rating

Using the information you've gathered about the performance appraisal process, I'm now going to show you 3 more interview segments so that you can become more acquainted with and practice some of the definitions and concepts we just talked about. Again, you will be rating the performance of the interviewer, the man with glasses seated behind the desk. In addition to coding your ratings on your opscan sheet, write out justifications for your evaluation of each of the 3 segments on the sheet provided. This will prove beneficial when we go back over your ratings together.

Have subjects write first name and Soc. Sec number on justification sheet.

Distribute numbered justification sheet for practice.

Show three segments, stopping after each for rating.
Seg 1, lines 33-39; seg 2, lines 40-46; seg 3, lines 47-53.

Feedback and Discussion of Practice Rating

When group is done, have them write their first name in pencil on top line of blue opscan.

Then collect opscans and give the group 10 min break.

Write ratings from opscans on board, ratee by dimension for each rater.
Hand back opscans.

These are your ratings of the 3 interview segments. You can see that some of you have avoided making the 2 rating errors to a better extent than others. Our goal is to make sure that you all are avoiding halo and leniency error. Let's take a look at each of your ratings individually.

For each rater:

Ask rater to read or describe their reasons for the way in which they rated each of the 3 segments.
Ask group how on target that rater was in avoiding halo and leniency in each of the 3 segments.
As the trainer, make sure each specific rating is discussed in the extent to which the rater avoided halo and leniency error.

Any questions concerning the definition of halo error?
Any questions concerning the definition of leniency error?
Any questions concerning the relation between your ratings and the amount of halo error in them?
Any questions concerning the relation between your ratings and the amount of leniency error in them?

Summary of Training

Soon, you will have one more opportunity to rate the performance of the interviewer. At this point, I would like to summarize the main points in our discussion on improving the quality of performance ratings.

First of all, when a psychologist talks about halo error or bias, to what are they referring? (central impression, rating someone similarly on all dimensions)

And for one final time, what is leniency error or bias? (when a rater consistently rates people very favorably)

How did you all learn more about these biases? (I talked about definitions of the errors, discussed everyday examples, I showed you some job performance examples, you practiced rating and avoiding these errors, you got feedback on how well you avoided the rating errors)

What did you get out of practicing, and then receiving feedback? (knowing how to avoid making these rating errors)

And now the most important question-- how can you improve the quality of your ratings? (observe all job relevant behavior and use the job dimensions effectively, avoid errors by including both high and low ratings for each performance segment, and keep ratings low)

At this point, you all can take a 10 minute break. When you come back, you will give one last set of interviewer performance ratings, and then fill out a couple of surveys.

Posttest

Show three performance segments, stopping after each for rating.
Seg 1, lines 54-60; seg 2, lines 61-67; seg 3, lines 68-74.

Then read instructions for aptitude test.
Then tell subjects to complete aptitude retest, lines 75-174.

Then have subjects respond to the training information survey,
lines 175-194.

Sign extra credit materials, have intro. students fill out 4 orange opscans, with their name, coded ID#, and in seat number code 002.
Give them debriefing about purpose of the experiment and answer any questions.

Appendix F

Control Training Protocol

Materials
dress in semi-formal clothes (look professional)
training protocol for condition assigned to that session
overview of training conditions
VCR (monitor will be in rooms already)
connection cable for VCR to monitor
chalk
extra coding pencils
training packet (interviewing survey, rating forms, aptitude retest, training information survey)
orange opscans (enough for 4 opscans per intro. student)
rating justification sheets
blue data opscans
informed consent forms

using 2nd half of the training videotape-- larger VCR start:
smaller VCR start:

***Approximate the time durations on the training overview as you go along and conduct the session.

Introduction and Pretest

Distribute informed consent forms.

Thank you all for coming. This important study is investigating the ways in which we can train people to evaluate job performance. Today's experimental session will run about 3 hours in total. Because of the length of this session, I will be letting you all take breaks at certain points in the middle of the experiment. First, I'd like to hand out some of the materials will be using in this session.

Distribute training packet and blue opscan form.
Have subjects code their Soc Sec. number in lines 1-9,
code their condition on line 10 (FOR=1, RET=2, CON=3)
code responses to interviewing survey in columns A-Y in top box.

At this point, I would like you all to view 3 performance segments on the VCR. After each segment, I will give you time to make performance ratings. You will be evaluating the interviewer. He is seated behind the desk and is wearing glasses.

Show three segments, stop after each for rating.
Tell subjects to skip line 11.
Seg 1, lines 12-18; Seg 2, lines 19-25, Seg 3, lines 26-32.

Perf. Appraisal, Multidimensionality of Work, Beh. Observation Lecture

You all just completed 3 performance appraisals after watching 3 interview segments on the VCR. Some of you may have had experience in your past with this type of task. Managers and supervisors commonly evaluate the performance of their subordinates to insure that their group or organization is running the best way possible. Often, this information on the quality of an employee's performance is presented, in turn, to the employee. This feedback occurs so that the employee can become more aware of the ways in which they are performing well, and the ways in which that employee may not be performing well. This performance information that is presented to the employee (as you might imagine) is extremely important to the overall performance of the organization.

On the performance appraisal forms that you completed earlier, you might also remember that there were a number of performance dimensions which were to be used in rating the interviewer. You all, in essence, made an employee performance rating for each of these performance dimensions. That is, for each employee doing a particular job (in this case, the interviewer in each of the 3 segments) you made a handful of ratings. For each interview segment, you made 6 or 7 performance ratings. Well, why did you make all these ratings for each segment? Shouldn't an overall performance rating be enough? The answer is that one overall rating is probably not enough to capture the quality of performance of that employee. Most jobs are not unidimensional. An employee could be strong on certain rating dimensions, but poor on others. And as you might conclude, all this information is very important to the organization and to the feedback to the employee. Work performance is viewed largely as being multidimensional in nature. And this is why it is common to see quite a few rating dimensions on performance appraisal forms.

However, a good performance appraisal system cannot be built solely upon a rating form which sufficiently describes a job along a handful of rating dimensions. In order for performance ratings to give an accurate portrayal of an employee's performance, the rater must pay close attention to that employee and that employee's work performance. And because this work performance is variable and multidimensional, the person who is making the ratings needs to attend to as much of the employee's specific work behaviors as is possible. That is, in order for a supervisor to be a good rater of employee performance, he or she must be an attentive observer of what that employee does and says in the line of work. Being a good observer in making performance ratings for each job dimension is the first step towards accurate and useful performance evaluations.

General Job Behaviors and Dimensions Discussion

Now that I have talked about the general nature of the performance appraisal process, I'd now like to talk about the specific job which we have been concerned with, and will continue to focus upon. The job is an interviewer of job applicants at a mid-sized, technologically-oriented organization. Thinking about what you observed in those 3 interview segments, and using the knowledge and experience you possess about such interviewers of job applicants, what behaviors are commonly required in such a job?

Examples are below.

1. Communicate with selection people, to select potential interviewees, help in making selection decisions.

2. Contact interviewee and set up interview.
3. Gather knowledge about specific open position.
4. Set up meetings with past position holders, potential supervisors, potential coworkers.
5. Organize interview, write questions concerning education, training, experience.
6. Establish rapport with interviewee.
7. Give preview of the job, organization.
8. Answer interviewees' questions.

Now that we've identified a lot of the behaviors required of an interviewer of job applicants, let's go back to the original rating dimensions which you used. Most rating forms, like the ones you had in front of you, are put together after someone has taken into account behaviors required on the job. In constructing the rating dimensions, the job behaviors are basically grouped into major categories. The rating dimensions for the job of interviewer are:

Write general names on board.

1. Rapport building- The extent to which the interviewer put the applicant at ease, built trust, and was socially appropriate.
2. Organization of the interview- The extent to which the interview had a clear format which was followed.
3. Questioning skill- The extent to which questions were tactfully presented.
4. Relevance of questions- How useful the interviewer's questions were for assessing the knowledge, skills, abilities, and characteristics of the applicant.
5. Company and job preview- The extent to which knowledge concerning the company and the job is appropriately conveyed.
6. Answering the applicant- The demonstrated willingness for answering the applicant's questions and the extent towhich the responses provided meaningful information.
7. Overall evaluation- Your general impression of the interviewer's overall performance in the interview.

Do you have any questions about the definitions of each of the rating dimensions?

Organizational Psychology Lecture/ Discussion

I would like to start off this portion of the session by talking for 10 minutes or so about the general concept of job satisfaction. Industrial/ Organizational psychologists, (the researchers who study human behavior in the work setting), have long been concerned with job satisfaction and how it relates to what people do on the job. Those of you who have had jobs may also have had experience with what job satisfaction is, and how its presence or absence might affect employees.

There are probably 3 reasons as to why psychologists have been so interested in job satisfaction. One is cultural, in that in the United States we value individual freedom, personal growth, and opportunity. In the fabric of this nation is the implicit belief that everyone has a right to a satisfying and rewarding job. The second reason for interest is functional in nature.

Research has shown that satisfaction is related to important variables like employee absenteeism, turnover, and performance. Though we're not sure whether job satisfaction has a causal relation to these variables, we do know that levels of job satisfaction are associated with certain levels of these variables. And in virtually all organizations, it is a chief objective to try and reduce absenteeism, reduce turnover (employees leaving the organization), and we also try and increase the quality of performance. Increasing levels of job satisfaction may accomplish these desired ends. The third reason for the study of satisfaction is that research in the area has been proceeding for a long time based on early landmark studies. Ground-breaking work in the 1920's revealed that employees had suprisingly strong feelings and opinions about their work. These studies dramatically shifted the work variables studied by psychologists. Economic and structural factors became less important, and interpersonal and attitude factors began to be emphasized.

As a result of this research, scientists have come to somewhat agree on a general definition of job satisfaction. They define job satisfaction as a pleasurable or positive emotional state resulting from the appraisal of one's job or job experiences. Unlike morale, which is a group response, job satisfaction is an individual response. The morale of a group could be high, while a person within that group could be dissatisfied with his or her job. The reverse could also be true.

It was initially thought that employees could have an overall feeling of liking for a job ranging from from very low to very high. This was known as global job satisfaction. Researchers have since learned that many factors contribute to overall feelings about jobs. Thus, 2 people could feel the same level of global job satisfaction, but for different reasons. Because of this, psychologists began examining specific facets or components of job satisfaction. Based on this perspective, researchers began measuring how people feel about various aspects of a job.

Several theories have been proposed to explain why people are satisfied (or dissatisfied) with their jobs. However, no one theory has been able to convincingly explain satisfaction in terms of experimental data. This again suggests that job satisfaction is a complex phenomenon with many causes, and that no one theory has been successful in incorporating all these causes. Each theory seems to explain a piece of the puzzle, but a complete and all-encompassing understanding is beyond its scope.

According to personal need and value-based theories, satisfaction is the extent to which a job meets or fulfills values. This is a process involving just the individual. Social comparison theories postulate that satisfaction is derived from a comparison between the self and others in similar jobs. A theory called the opponent process theory puts forth that satisfaction is primarily physiological. The central nervous system is responsible for satisfaction, especially in regard to protecting a person from extreme emotions. Finally, the 2-factor theory posits that work conditions are the sources of satisfaction.

All in all, each of these theories has, in its own way, contributed to our understanding of job satisfaction. It seems unlikely , however, that researchers will develop one general theory of job satisfaction.

As I have mentioned, job satisfaction in itself has been a major variable in psychological research in industry and organizations. Other variables which I mentioned earlier are also very important to the operation of a company. Further, many researchers feel that these variables are directly related to job satisfaction. These variables are employee absenteeism, turnover, and performance.

In terms of absenteeism, many people tend to feel that employees who do not like their jobs are absent more often. However, research has shown that this may be a weak effect at best. One researcher has suggested that satisfaction measures can predict attendance only when attending is a result of special effort on the part of the employee. So, for example, satisfaction may predict who shows up for work in the middle of a raging snowstorm. Similarly, it has been concluded that if a company simply allows frequent absences (for example, lots of sick days and excused absences), employees will take advantage of that effortless situation regardless of their feelings of job satisfaction. In other words, if few rewards or sanctions were tied to abseteeism, probably no relationship would exist between satisfaction and absence. If employees are neither rewarded or punished for absence, their feelings about their jobs may be independent of their attendance. Other models linking satisfaction and attendance behavior cite specific important factors. These include: economic factors, work group norms, incentive systems, family responsibilities, and transportation problems. These factors help in showing that the connection between satisfaction and attendance is neither simple nor direct.

When talking about employee turnover, it does appear that the more people dislike their jobs, the more likely they are to quit. One researcher proposed that there are links between satisfaction and quitting. These links included thinking about quitting, looking for another job, intending to quit, and deciding to quit. The researcher contended that feelings of disstisfaction provoke thoughts of quitting, which in turn prompt the search for another job. This model was a step forward in thinking of the process from job dissatisfaction to turnover instead of repeatedly attempting to investigate the direct relationship between satisfaction and turnover.

The relation between satisfaction and employee performance has generated a great deal of interest in Industrial/Organizational psychology research. This is probably because we all would like employees to be both productive and happy with their work. Yet early research led many to the conclusion that at the most, satisfaction and performance were only slightly related. A controversy then arose as to whether satisfaction causes performance or performance causes satisfaction. Currently, most psychologists tend to agree that performance causes satisfaction. People get pleasure from their work after finding they are good at it. And managers should base rewards on past performance in the belief that this will reinforce desired performance. Other studies have investigated the conditions under which satisfaction and performance are related. It was found that the relation is stronger when rewards are based on performance. Therefore, people whose pay is based on performance, such as salespeople on commission, should be more satisfied with their specific performance than others paid on an hourly wage. In general, the satisfaction-performance relationship is not very large and is not consistent across different jobs.

What do you think makes you satisfied when you're at work?

What do you think-- is a happy worker more productive? Personal examples?

Write examples on board.

Soon, you will rate the interviewer's job performance again. But first, here are more examples of the interviewer's job behaviors.

Show six segments (1, 7, 16, 9, 12, 18).

Practice Rating

Using the information you've gathered earlier about the performance appraisal process, I'm now going to show you 3 more interview segments so that you can become more acquainted with and practice some of the definitions and concepts we talked about. Again, you will be rating the performance of the interviewer, the man with glasses seated behind the desk. In addition to coding your ratings on your opscan sheet, write out justifications for your evaluation of each of the 3 segments on the sheet provided. This will prove beneficial when we go back over your ratings together.

Have subjects write first name and Soc. Sec number on justification sheet.

Distribute numbered justification sheet for practice.

Show three segments, stopping after each for rating.
Seg 1, lines 33-39; seg 2, lines 40-46; seg 3, lines 47-53.

Feedback and Discussion of Practice Rating

When group is done, have them write their first name in pencil on top line of blue opscan.

Then collect opscans and give the group 10 min break.

Write ratings from opscans on board, ratee by dimension for each rater.
Hand back opscans.

These are your ratings of the 3 interview segments. You can see that there are some similarities, and some differences in the group's ratings. Let's take a look at each of your ratings individually.

For each rater:

Ask rater "What is your general impression of your ratings?", "How did you rate?"
Ask group "In general, how on target are these ratings?".
As the trainer, make a general comment as to the ratings' level of similarity to the other trainees' ratings.

Any questions?

Organizational Psychology Discussion

At this point, I would like to switch gears. Before, I lectured for awhile about job satisfaction. Now, I'd like to have more of a group discussion about job satisfaction, and hear more of what you have to say on the subject.

Do you think that every employee has the right to be satisfied with their work? What is the role of the manager in facilitating employee satisfaction? How about the president of the company's role?

Does feedback on level of performance lead to levels of satisfaction, or does satisfaction lead to performance?

In determining satisfaction, what do you think is more important--internal values or social comparison?

At this point, you all can take a 10 minute break. When you come back, you will give one last set of interviewer performance ratings, and then fill out a couple of surveys.

Posttest

Show three performance segments, stopping after each for rating.
Seg 1, lines 54-60; seg 2, lines 61-67; seg 3, lines 68-74.

Then read instructions for aptitude test.
Then tell subjects to complete aptitude retest, lines 75-174.

Then have subjects respond to the training information survey, lines 175-194.

Sign extra credit materials, have intro. students fill out 4 orange opscans, with their name, coded ID#, and in seat number code 002.
Give them debriefing about purpose of the experiment and answer any questions.

Appendix G

Interviewing Test

(bubble your social security number in lines 1-9 on your opscan.)
(bubble the number the experimenter dictates to you in line 10.)

Scale 1. Multiple Choice

Please record the best answers in the top box of your opscan sheet.

A) For the most part, the ability of interviewers to identify successful employees is:
1. on the high side.
2. about average for applicant screening methods.
3. on the low side.
4. too variable to draw conclusions.

B) In terms of effective interviewing skills for the applicant, the most important non-verbal behavior is:
1. smiling.
2. nodding your head.
3. body position.
4. eye contact.

C) In terms of accurately assessing applicants, the worst interviewing strategy is for the interviewer to:
1. ask an identical set of questions to each applicant.
2. cause the applicant to feel stress while responding.
3. ask different questions to each applicant.
4. have several people judge the applicant's performance.

D) As an applicant, if you have to say something negative about yourself it is best to do it:
1. early in the interview.
2. in the middle of the interview.
3. late in the interview.

E) Typically, the first goal of an interviewer is to:
1. force the applicant to tell the truth.
2. sell the organization to the applicant.
3. put the applicant at ease.
4. find out the applicant's motivation level.

F) The typical interviewer makes his/her decision about an applicant:
1. early in the interview.
2. in the middle of the interview.
3. late in the interview.
4. after the interview is over.

G) Which of the following pieces of information would have the most influence on the typical interviewer?
1. graduating from college with honors.
2. receiving a service award in connection with your job.
3. having been fired once.
4. a history of perfect work attendance.

H) The information most likely to lead to accurate decisions on the part of an interviewer are questions:
1. about educational backround.
2. related to performing the job in question.
3. concerning hobbies that the applicant is "into".
4. about general work experience.

I) As an applicant, the strategy most likely to increase your chance of getting the job is to:
1. research the organization prior to your interview.
2. dress in appropriate business attire.
3. be on time for your interview.
4. use a firm handshake.

J) When describing the job to the applicant, the interviewer should:
1. make the job appear more attractive than it really is so that the applicant wants the job.
2. make the job appear less attractive than it really is so that the interviewer is sure the applicant wants the job.
3. attempt to describe the job as accurately as possible so that the applicant can make an informed decision.

K) Which of the following situations would be the most desirable for a marginally-qualified applicant?
1. The interviewer asks the same questions to all applicants.
2. The interviewer does most of the talking during the interview.
3. The marginal applicant's interview follows the interview of a well-qualified applicant.
4. The interviewer asks many questions about doing the job in question.

L) All other things being equal, it is best for the applicant to be:
1. similar to the interviewer.
2. taller than the interviewer.
3. of the opposite sex of the interviewer.
4. more attractive than the interviewer.

M) Which of the following statements is the most accurate?
1. As an interviewer gains more interviewing experience, his/her accuracy improves.
2. Men are better interviewers than women.
3. Interviewers are more accurate after being trained to conduct interviews.

N) The way you dress for an interview is important because:
1. your dress is an indication of your personality.
2. interviewers are influenced by their first impression.
3. it reflects how you will dress on the job.
4. your effort to look good reflects how badly you want the job.

O) As an applicant, the best way to ensure that a job suits your needs is to:
1. keep an open mind about each job for which you interview.
2. always use a recruiting service.
3. apply to jobs for which you are marginally qualified.
4. prepare a list of questions to ask each interviewer.

Scale 2. Interviewing Experience

P) How many formal interviews (e.g., a job interview) have you participated as an interviewee?

0	1	2	3	4	5
none	1-2	3-4	5-6	7-8	9 or more

Q) How many formal interviews (e.g., a job interview) have you participated as an interviewer?

0	1	2	3	4	5
none	1-2	3-4	5-6	7-8	9 or more

R) Have you taken any classes or received any training that included interviewing techniques?

0	1	2	3
none	a little amount 0 - 1 hr.	a moderate amount 1 - 2 hrs.	a large amount > 2 hrs.

S) How many magazine articles, journal articles, book sections, or pamphlets/brochures have you read which focused on interviewing techniques?

0	1	2	3	4	5
none	1-2	3-4	5-6	7-8	9 or more

T) How confident are you in your ability to rate a job applicant?

0	1	2	3	4
not at all confident		moderately confident		very confident

U) Rate your ability to judge a job interview as a success or failure.

0	1	2	3	4
low ability		moderate ability		high ability

V) What is your knowledge of the "personal" kinds of questions asked in an interview?

0	1	2	3	4
no knowledge		moderate knowledge		a lot of knowledge

W) What is your knowledge of the "past work experience" questions asked in an interview?

0	1	2	3	4
no knowledge		moderate knowledge		a lot of knowledge

X) What is your knowledge of the "educational backround" questions asked during an interview?

0	1	2	3	4
no knowledge		moderate knowledge		a lot of knowledge

Y) How well do you understand the general structure of a job interview?

0	1	2	3	4
no understanding		moderate understanding		a lot of understanding

Rating Aptitude Test (PRF)

You are asked to answer the following questionnaire
by indicating the likelihood that an INTERVIEWER OF JOB APPLICANTS would agree or mark that particular item as true.

You are NOT being asked to respond regarding your personal level of agreement to these items.

To repeat: PLEASE ANSWER THESE ITEMS THE WAY YOU FEEL AN INTERVIEWER OF JOB APPLICANTS WOULD RESPOND.

Indicate your response to each of the questionnaire items by coding 1 through 9 on the blue opscan sheet provided by the experimenter.

Interviewer of job applicants:

1	2	3	4	5	6	7	8	9
definitely will NOT say true		probably will NOT say true		maybe say true		probably will say true		definitely will say true

75. People should be more involved with their work.

76. I find that I can think better when I have the advice of others.

77. I very seldom make careful plans.

78. I don't have the staying power to do work that must be very accurate.

79. Often I stop in the middle of one activity in order to start something else.

80. I spend quite a lot of time keeping my belongings in order.

81. People consider me a serious, reserved person.

82. The motion of water in a river can almost hypnotize me.

83. If I feel sick, I don't like to have friends or relatives fuss over me.

84. There are many activities that I prefer to reading.

85. I seldom set standards which are difficult for me to reach.

86. I delight in feeling unattached.

87. When I go on a trip I prepare a timetable beforehand.

88. When I hit a snag in what I am doing, I don't stop until I have found a way to get around it.

89. I am careful to consider all sides of an issue before taking action.

90. I feel comfortable in a somewhat disorganized room.

91. I spend a good deal of my time just having fun.

92. I rarely notice the texture of a piece of clothing.

93. I would like to be married to a protective and sympathetic person.

94. I like to read several books on one topic at the same time.

95. I enjoy difficult work.

Interviewer of job applicants:

1	2	3	4	5	6	7	8	9
definitely will NOT say true		probably will NOT say true		maybe say true		probably will say true		definitely will say true

96. Family obligations make me feel important.

97. I like to be with people who change their minds often.

98. If I run into great difficulties on a project, I usually stop work rather than try to solve them.

99. I often say the first thing that comes into my head.

100. When writing something, I keep my pencils sharpened.

101. Most of my friends are serious-minded people.

102. I like to feel sculptured objects.

103. I prefer not being dependent on anyone for assistance.

104. I would rather work in business than in science.

105. I have rarely done extra studying in connection with my work.

106. People who try to regulate my conduct with rules are a bother.

107. Before I ask a question, I decide exactly what it is I need to find out.

108. I am willing to work longer at a project than are most people.

109. I am pretty cautious.

110. I am often disorganized.

111. At times I get fascinated by some unimportant game and play with it for hours.

112. I have never seen a statue that reminded me of a real person.

113. I try to share my burdens with someone who can help me.

114. I am more at home in an intellectual discussion than in a discussion of sports.

115. I will not be satisfied until I am the best in my field of work.

Interviewer of job applicants:

1	2	3	4	5	6	7	8	9
definitely will NOT say true		probably will NOT say true		maybe say true		probably will say true		definitely will say true

116. I would feel lost and lonely roaming around the world alone.

117. I tend to start right in on a new task without thinking about the best way to do it.

118. If I get tired while playing a game, I generally stop playing.

119. When I go to the store, I often come home with things I had not intended to buy.

120. A place for everything and everything in its place is the way I like to live.

121. I would prefer a quiet evening with friends to a loud party.

122. Sometimes I feel like stepping into mud and letting it ooze between my toes.

123. The person I marry won't have to spend much time taking care of me.

124. I tend to shy away from intellectual discussions.

125. I try to work just hard enough to get by.

126. I could live alone and enjoy it.

127. Often when I telephone someone, I make a list of things to discuss.

128. I have spent hours looking for something I needed to complete a project.

129. Rarely, if ever, do I do anything reckless.

130. I often forget to put things back in their places.

131. Most of my spare moments are spent relaxing and amusing myself.

132. I don't care whether I drink water from a fine glass or from a paper cup.

133. I want to be sure someone will take care of me when I am old.

134. I like magazines offering thoughtful discussions of politics and art.

135. I would work just as hard whether or not I had to earn a living.

136. I respect rules because they guide me.

Interviewer of job applicants:

1	2	3	4	5	6	7	8	9
definitely will NOT say true		probably will NOT say true		maybe say true		probably will say true		definitely will say true

137. I rarely consider the daily weather report when deciding what to wear.

138. I don't believe in sticking to something when there is little chance of success.

139. Many of my actions seem to be hasty.

140. If I have to pack a suitcase, I usually organize it very well.

141. Even if I had the money and the time, I wouldn't feel right just playing around.

142. One of my favorite pastimes is sitting before a crackling fire.

143. I usually make decisions without consulting others.

144. Serious books are of little use to me.

145. I do not let my work get in the way of what I really want to do.

146. I would not mind living in a very lonely place.

147. When I make something I want to know exactly what it will look like when finished.

148. If I want to know the answer to a question, I sometimes look for it for days.

149. Emotion seldom causes me to act without thinking.

150. I have a lot of trouble keeping an accurate record of my expenses.

151. Rarely, if ever, do I turn down a chance to have a good time.

152. I don't get any particular enjoyment from sitting in the sun.

153. I like to ask other people's opinions concerning my problems.

154. I think I would enjoy studying most of my life so I could learn as many things as possible.

155. My goal is to do at least a little bit more than anyone else has done before.

Interviewer of job applicants:

1	2	3	4	5	6	7	8	9
definitely will NOT say true		probably will NOT say true		maybe say true		probably will say true		definitely will say true

156. Adventures where I am on my own are a little frightening to me.

157. I live from day to day without trying to fit my activities into a pattern.

158. If I become tired I set my work aside until I am more rested.

159. I have often broken things because of carelessness.

160. My work is always well organized.

161. I only celebrate very special events.

162. Certain pieces of music remind me of pictures or moving patterns of color.

163. I prefer to face my problems by myself.

164. I really don't know what is involved in any of the latest cultural developments.

165. In my work I seldom do more than is necessary.

166. I would like to be alone and my own boss.

167. I try to plan my future so that I can tell what I will be doing at any given time.

168. I rarely let anything keep me from an important job.

169. I have a reserved and cautious attitude toward life.

170. I rarely clean out my bureau drawers.

171. I pride myself on being able to see the funny side of every situation.

172. I don't get any particular enjoyment from having my neck massaged.

173. If I ever think that I am in danger, my first reaction is to look for help from someone.

174. I do almost as much reading on my own as I did for classes when I was in school.

Rater Error Training Test

Some of the material below may be somewhat unfamiliar to you. Please answer TRUE or FALSE to the best of your ability.

If TRUE, code bubble 1.
If FALSE, code bubble 2.

175. Halo error takes place only when the rater has positive feelings about a ratee.

176. Leniency error occurs when a rater's central impression causes a uniform distribution of ratings.

177. The best way to combat leniency error is for raters to include both high and low ratings in their assessment of a ratee.

178. The best way to combat halo error is for raters to make sure that raters are not too "easy going" or overly favorable in their evaluations of ratees.

179. A teacher gives a student the same low rating across all academic performance dimensions. This is an example of halo error.

180. A manager feels that their employee is the best in the work group on one particular rating dimension. Keeping this in mind, the manager rates that employee uniformly on the positive end of the rating scale across all of the job performance dimensions. This is an example of halo error.

181. An army sergeant rates his subordinates on several infantry performance dimensions. Most of the members of his platoon are rated favorably. This is an example of leniency error.

182.

Dimension 1	Dimension 2	Dimension 3	Dimension 4
6	6	6	6

On this rating scale of 1 to 7, a higher rating is a more favorable rating. The above rating distribution is an example of halo error.

183.

Dimension 1	Dimension 2	Dimension 3	Dimension 4
4	3	5	3

On this rating scale of 1 to 7, a higher rating is a more favorable rating. The above rating distribution is an example of leniency error.

184.

Dimension 1	Dimension 2	Dimension 3	Dimension 4
2	2	2	2

On this rating scale of 1 to 7, a higher rating is a more favorable rating. The above rating distribution is an example of halo error.

Frame-of-Reference Training Test

185. From a Frame of Reference training perspective, just knowing which behaviors belong to which performance dimensions is probably enough to improve the quality of a rater's ratings.

186. Frame of Reference training involves exposing a rater to a range of behaviors and the corresponding correct ratings for those behaviors.

187. Most often in Frame of Reference training, the rater is exposed to a range of behaviors which are concentrated on poor aspects of ratee performance.

188. The Frame of Reference trainer attempts to steer raters away from using standards in assessing a ratee's performance.

189. According to the Frame of Reference perspective, rater accuracy cannot be altered by the rater knowing what are good, average, and poor examples of job behavior within each performance dimension.

190. In Frame of Reference training, raters adjust their perspective of a particular job by comparing examples of their ratings to true or target performance ratings.

191. The general goal of Frame of Reference training involves a minimal focus on increasing rating accuracy among raters.

192. An example of good performance on the Rapport Building rating dimension might be the interviewer asking about an applicant's cough, talking about the weather, and talking about possible locations where the applicant might work.

193. Asking an applicant about their involvement in community activities is probably an example of average performance on the Relevance of Questions rating dimension.

194. The interviewer leaving the office in the middle of an interview, and then making a long pause before asking about movies the applicant had seen recently is an example of poor performance on the Organization of Interview rating dimension.

Table 1

Descriptive Statistics for Frame-of-Reference Training Subjects

	Normative N=20		Idiosyncratic N=21	
	M	SD	M	SD
Pretest Elevation	.349	.287	.433	.320
Pretest Differential Elevation	.703	.420	.796	.391
Pretest Stereotype Accuracy	.547	.166	.568	.233
Pretest Differential Accuracy	.796	.199	.666	.184
Posttest Elevation	.534	.376	.604	.440
Posttest Differential Elevation	.418	.217	.543	.262
Posttest Stereotype Accuracy	.518	.191	.405	.138
Posttest Differential Accuracy	.571	.103	.426	.103
Interviewing Test (2nd scale)	15.40	4.93	15.57	5.29
Rater Error Test	4.05	1.57	4.00	1.10
Frame-of-Reference Test	7.05	1.05	6.38	1.63

Table 2

Descriptive Statistics for Rater Error Training Subjects

	Normative N=22		Idiosyncratic N=23	
	M	SD	M	SD
Pretest Elevation	.544	.431	.506	.376
Pretest Differential Elevation	.586	.273	.734	.400
Pretest Stereotype Accuracy	.557	.258	.582	.160
Pretest Differential Accuracy	.736	.220	.733	.160
Posttest Elevation	.519	.410	.313	.266
Posttest Differential Elevation	.637	.286	.571	.249
Posttest Stereotype Accuracy	.725	.300	.697	.201
Posttest Differential Accuracy	.664	.181	.755	.218
Interviewing Test (2nd scale)	16.14	4.67	16.30	5.98
Rater Error Test	8.09	1.41	8.52	1.12
Frame-of-Reference Test	6.27	1.67	6.52	.79

Table 3

Descriptive Statistics for Control Training Subjects

	Normative N=21		Idiosyncratic N=22	
	M	SD	M	SD
Pretest Elevation	.593	.394	.515	.504
Pretest Differential Elevation	.695	.261	.661	.435
Pretest Stereotype Accuracy	.549	.208	.506	.189
Pretest Differential Accuracy	.856	.362	.651	.162
Posttest Elevation	.654	.509	.599	.426
Posttest Differential Elevation	.588	.300	.663	.282
Posttest Stereotype Accuracy	.729	.353	.543	.220
Posttest Differential Accuracy	.547	.165	.564	.224
Interviewing Test (2nd scale)	16.14	4.67	16.30	5.98
Rater Error Test	4.10	1.61	4.23	1.97
Frame-of-Reference Test	6.57	1.40	6.23	1.38

Table 4

Unadjusted and Adjusted Interrater Reliability Coefficients

Training	Rating Aptitude	Pretest		Posttest	
		unadjusted	adjusted	unadjusted	adjusted
Frame-of-Reference	Normative	.95	.66	.98	.85
	Idiosyncratic	.93	.54	.96	.72
Rater Error Training	Normative	.97	.76	.94	.59
	Idiosyncratic	.98	.77	.97	.72
Control Training	Normative	.95	.63	.96	.72
	Idiosyncratic	.96	.75	.95	.63

Table 5

Repeated Measures ANOVA Summary Table for Elevation

Between Subjects Effects			
Source	df	Mean Square	F Value
Training	2	.383	2.43
Aptitude	1	.106	.67
Training by Aptitude	2	.225	1.43
Error	123	.157	

Within Subjects Effects			
Source	df	Mean Square	F Value
Measurement Occasion	1	.114	.69
Measurement Occasion by Training	2	.462	2.80
Measurement Occasion by Aptitude	1	.049	.30
Measurement Occasion by Training by Aptitude	2	.057	.34
Error (Measurement Occasion)	123	.165	

Note. * $p < .05$, ** $p < .01$.

Table 6

Repeated Measures ANOVA Summary Table for Stereotype Accuracy

Between Subjects Effects			
Source	df	Mean Square	F Value
Training	2	.369	6.46**
Aptitude	1	.184	3.21
Training by Aptitude	2	.070	1.23
Error	123	.057	

Within Subjects Effects			
Source	df	Mean Square	F Value
Measurement Occasion	1	.185	4.13*
Measurement Occasion by Training	2	.352	7.84**
Measurement Occasion by Aptitude	1	.191	4.25*
Measurement Occasion by Training by Aptitude	2	.014	.30
Error (Measurement Occasion)	123	.045	

Note. * $p < .05$, ** $p < .01$.

Table 7

Repeated Measures ANOVA Summary Table for Differential Accuracy

Between Subjects Effects			
Source	df	Mean Square	F Value
Training	2	.265	5.49**
Aptitude	1	.230	4.77*
Training by Aptitude	2	.197	4.09*
Error	123	.048	

Within Subjects Effects			
Source	df	Mean Square	F Value
Measurement Occasion	1	1.402	43.08**
Measurement Occasion by Training	2	.271	8.31**
Measurement Occasion by Aptitude	1	.169	5.18*
Measurement Occasion by Training by Aptitude	2	.074	2.28
Error (Measurement Occasion)	123	.033	

Note. * $p < .05$, ** $p < .01$.

Table 8

Repeated Measures ANOVA Summary Table for Differential Elevation

Between Subjects Effects			
Source	df	Mean Square	F Value
Training	2	.026	.24
Aptitude	1	.198	1.80
Training by Aptitude	2	.044	.40
Error	123	.110	

Within Subjects Effects			
Source	df	Mean Square	F Value
Measurement Occasion	1	.972	9.83**
Measurement Occasion by Training	2	.321	3.24*
Measurement Occasion by Aptitude	1	.013	.13
Measurement Occasion by Training by Aptitude	2	.157	1.58
Error (Measurement Occasion)	123	.099	

Note. * $p < .05$, ** $p < .01$.

Table 9

Percent Change in Rating Aptitude Measured at PRF Retest

Qualitative Change in Raters' Responses to PRF	Training		
	FOR	RET	CON
Normative maintained Normative classification	60%	82%	48%
Normative decreased in threshold	30%	14%	33%
Normative decreased in sensitivity	5%	0%	5%
Normative changed to Idiosyncratic classification	5%	4%	14%
Idiosyncratic maintained Idiosyncratic classification	57%	35%	46%
Idiosyncratic increased in threshold	19%	30%	27%
Idiosyncratic increased in sensitivity	10%	0%	18%
Idiosyncratic changed to Normative classification	14%	35%	9%

Note. FOR = Frame-of-Reference Training, RET = Rater Error Training, CON = Control Training.

Figures

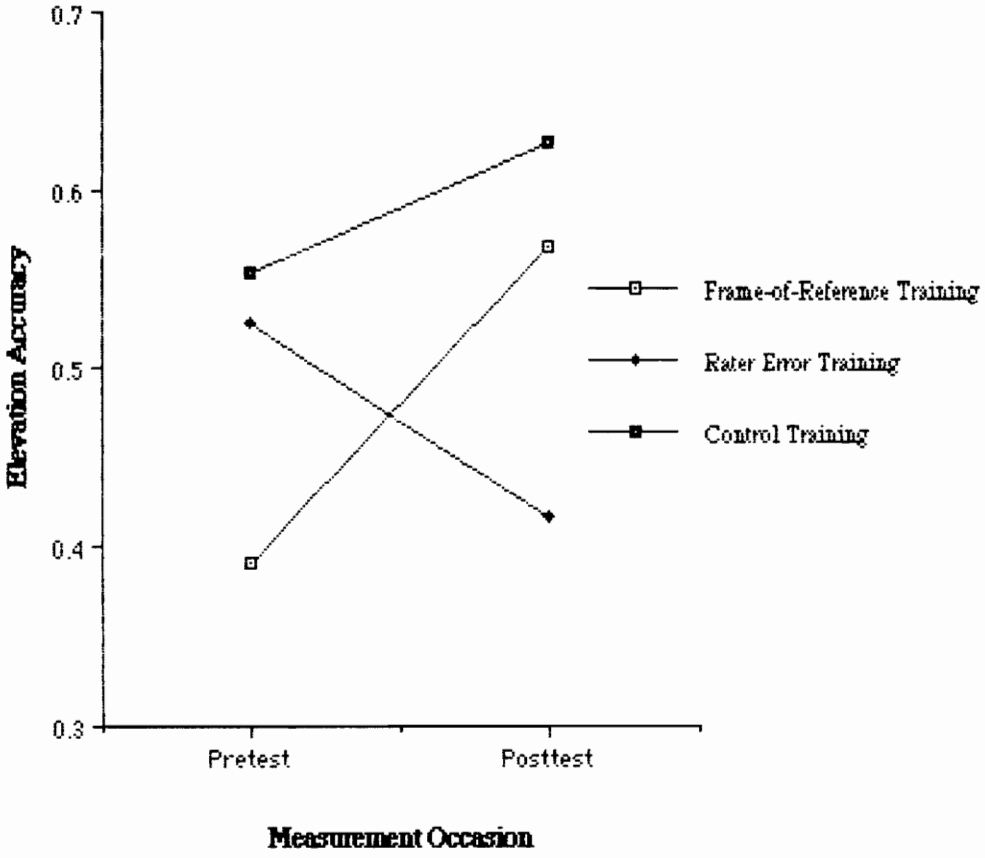


Figure 1

Rater Training Effects on Elevation

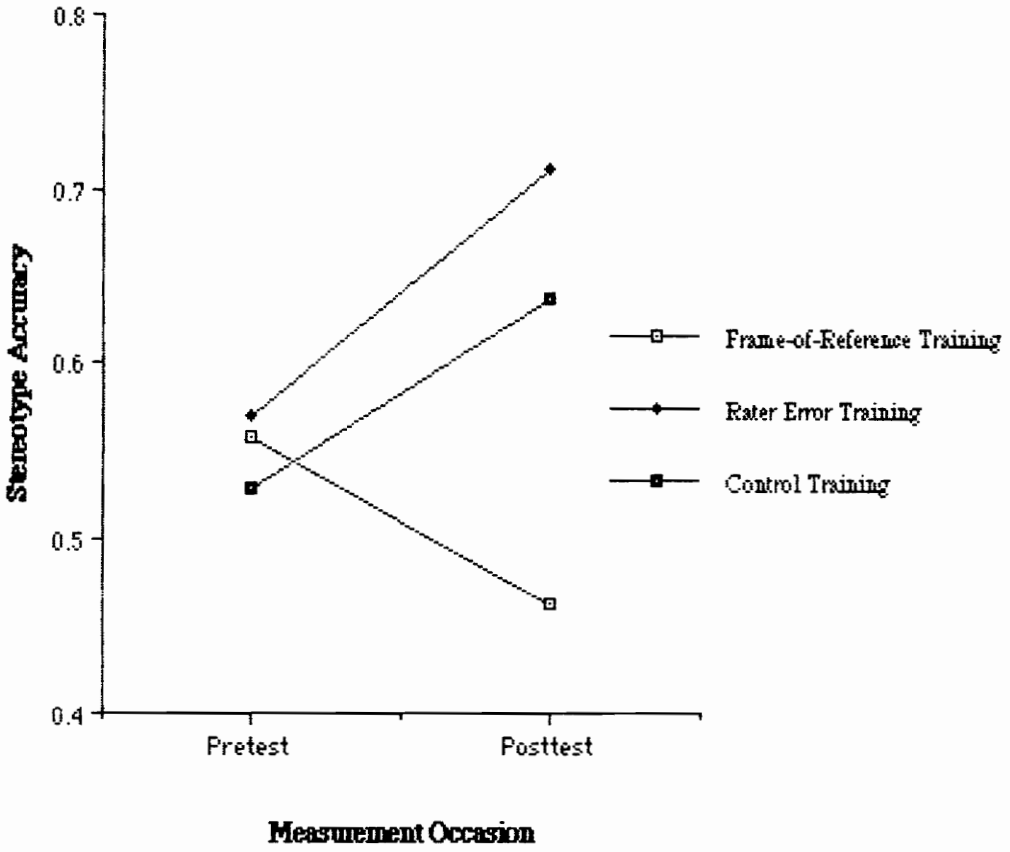


Figure 2

Rater Training Effects on Stereotype Accuracy

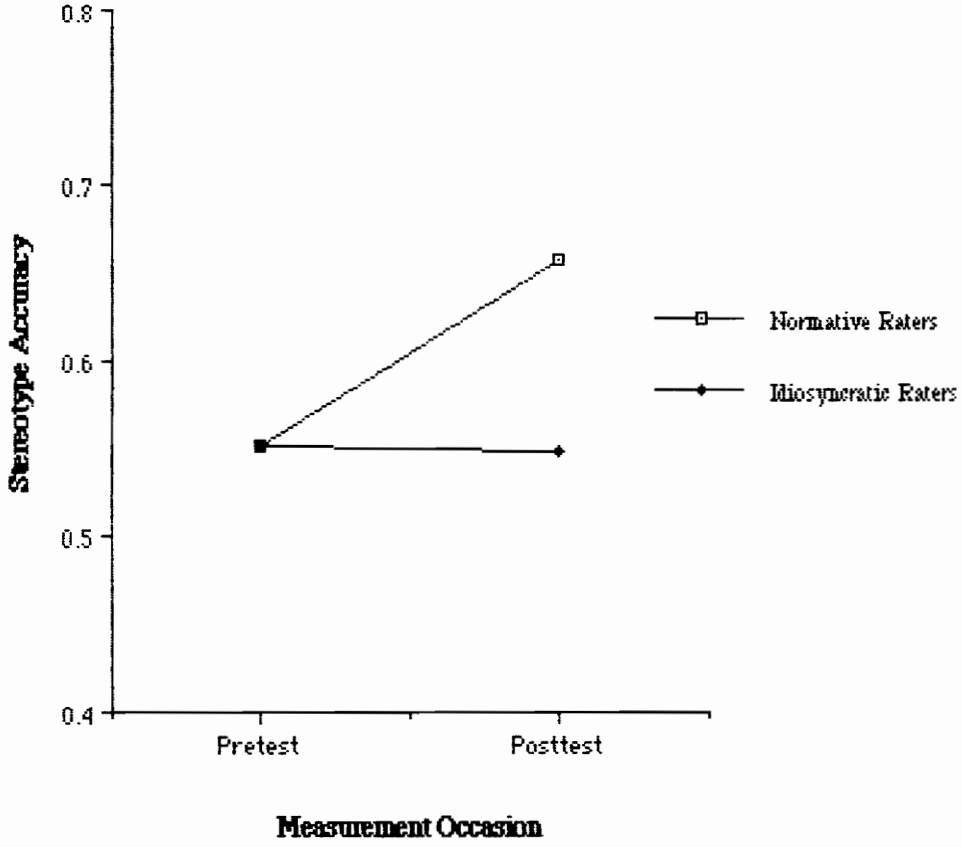


Figure 3

Pretest-Posttest Aptitude Differences in Stereotype Accuracy

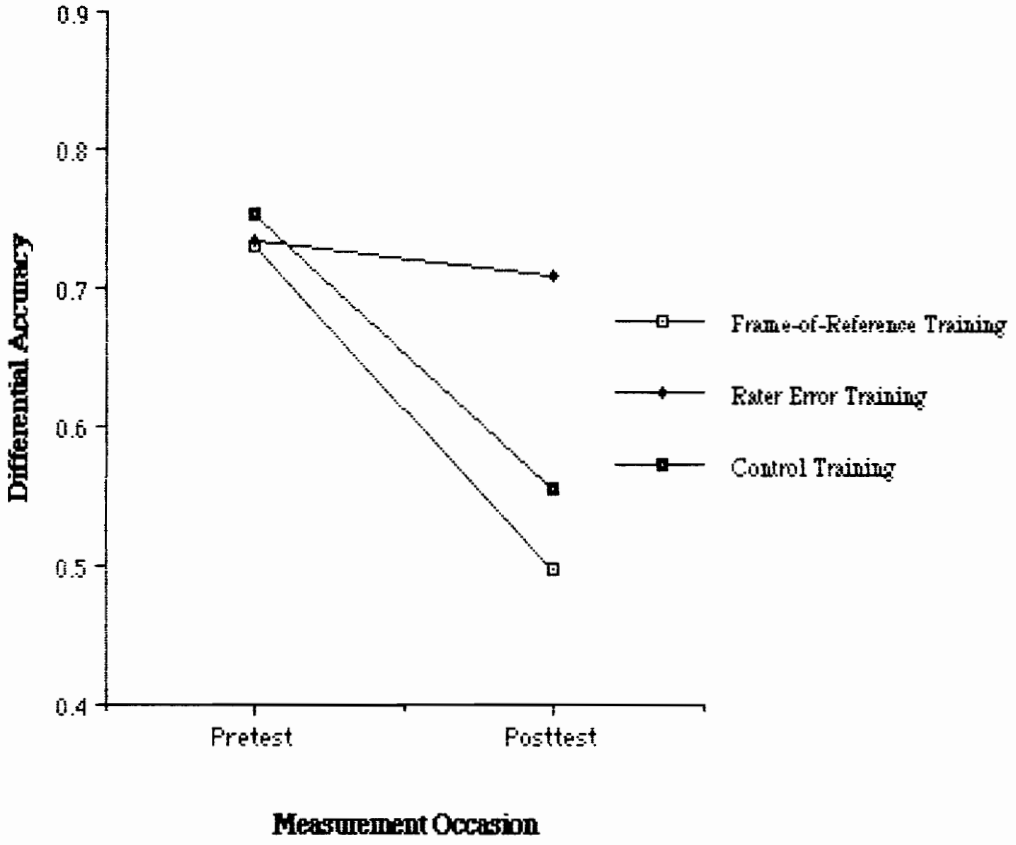


Figure 4

Rater Training Effects on Differential Accuracy

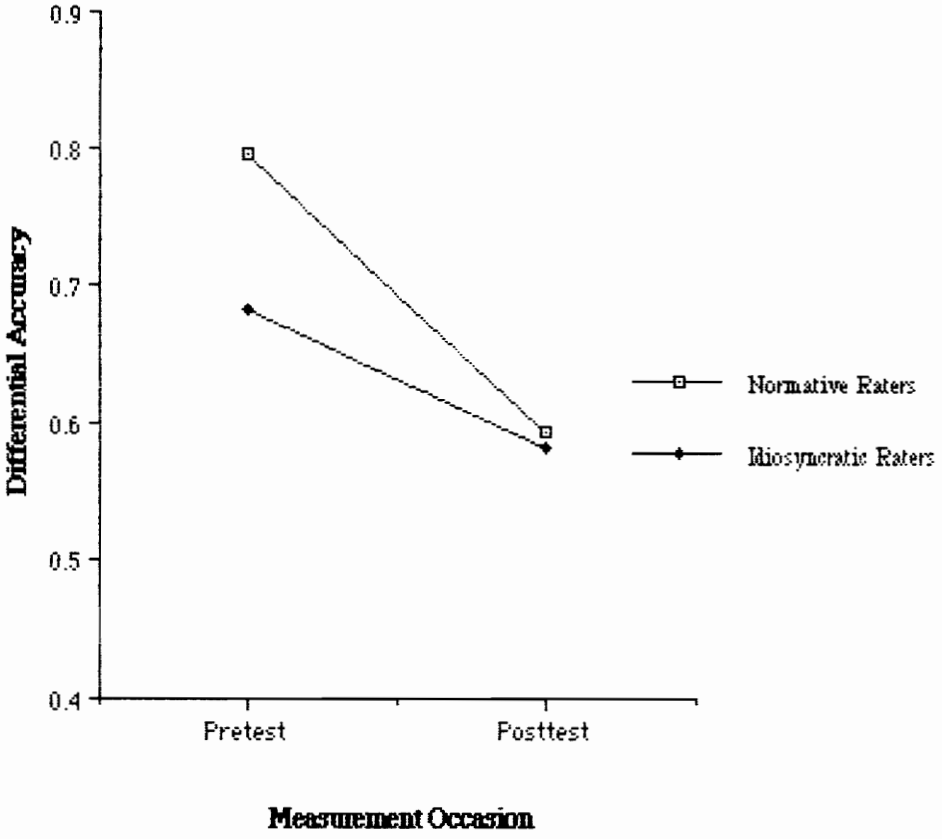


Figure 5

Pretest-Posttest Aptitude Differences in Differential Accuracy

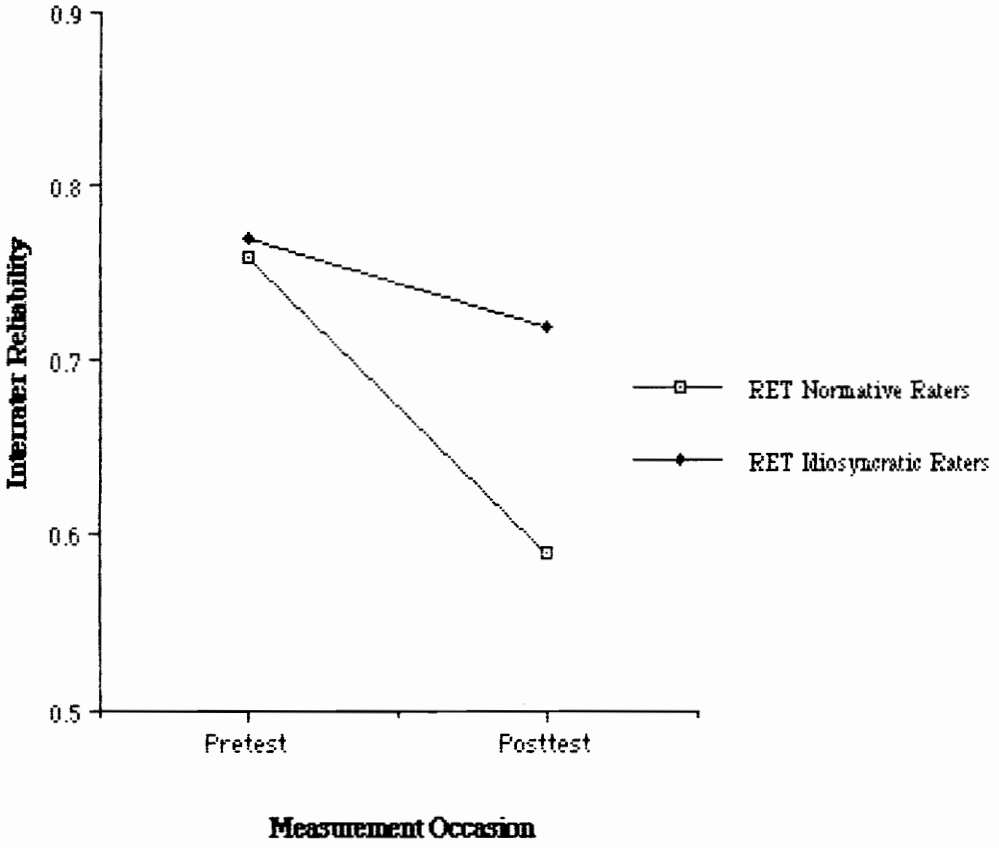


Figure 6

Rater Error Training Effects on Interrater Reliability

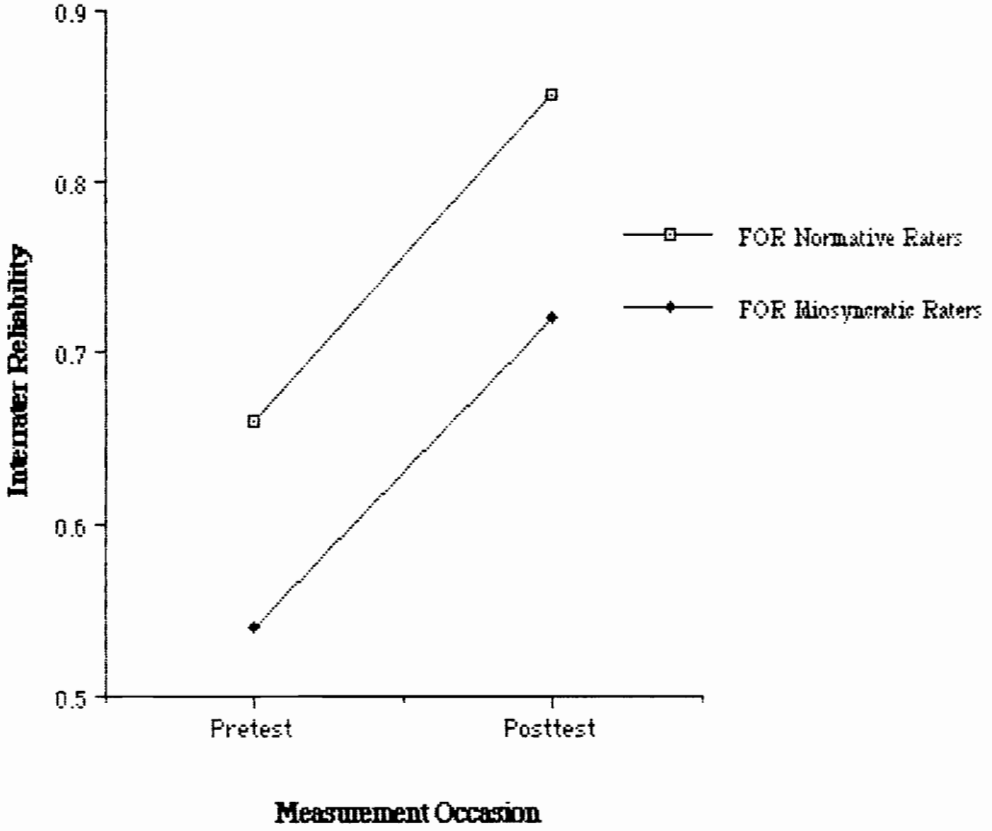


Figure 7

Frame-of-Reference Training Effects on Interrater Reliability

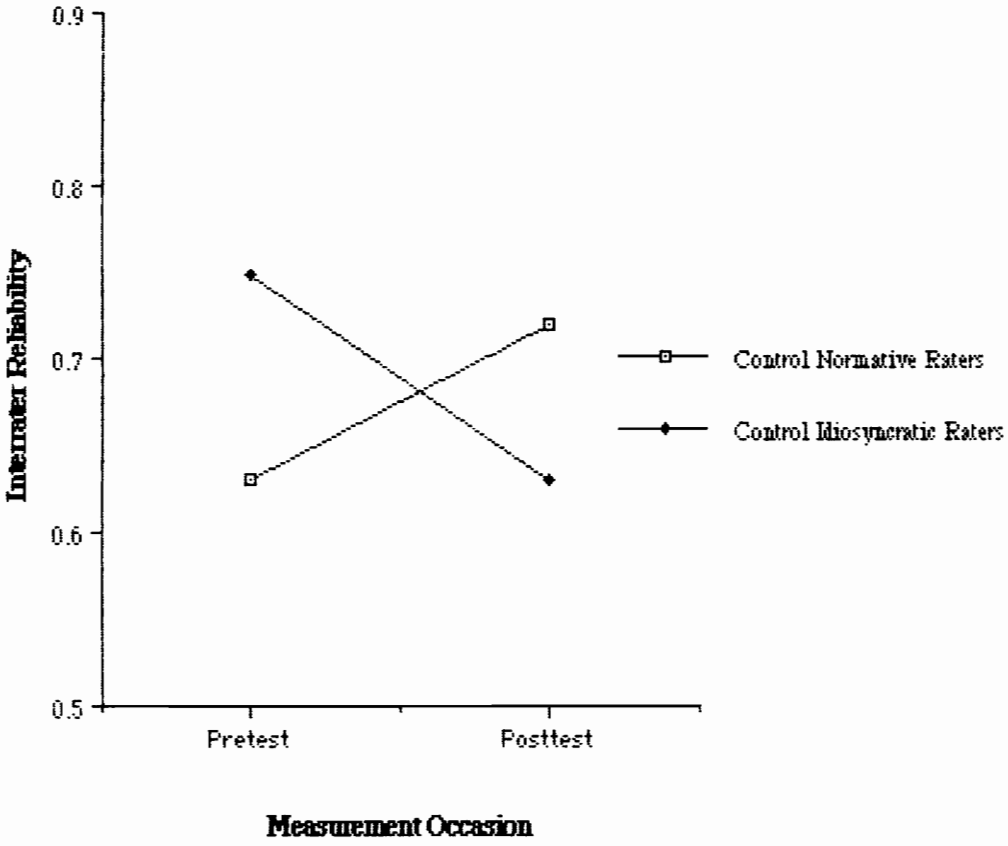


Figure 8

Control Training Effects on Interrater Reliability

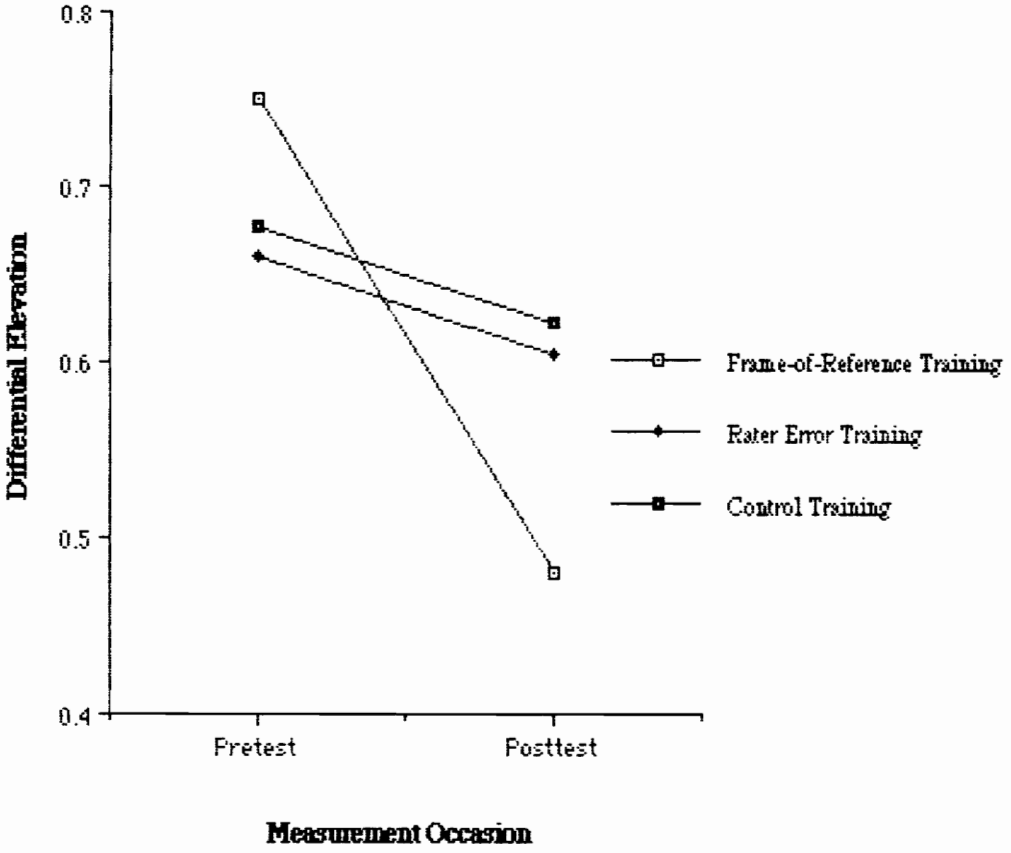


Figure 9

Rater Training Effects on Differential Elevation

Vita

DEAN THOMAS STAMOULIS

Date of Birth: 28 August 1966

Home Address and Phone
2415 Ridge Road
Blacksburg, Virginia 24060
(703) 552-1258

Office Address and Phone
Department of Psychology
Virginia Polytechnic Institute
and State University
(703) 231-6581

EDUCATION

1988-1990 **M.S. Industrial/ Organizational Psychology**
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
Blacksburg, Virginia

Thesis Title: The effects of frame-of-reference and rater error training on the accuracy of performance appraisals: Utilizing an aptitude-treatment approach.

1984-1988 **B.S. Psychology**
BROWN UNIVERSITY
Providence, Rhode Island

Thesis Title: Common problems and coping strategies reported by children.

PROFESSIONAL AND RESEARCH ACTIVITIES

1/91 - present Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, VA
Teaching Assistant for Industrial/ Organizational Psychology and Quantitative Topics undergraduate-level courses.

9/90 - present Neil Hauenstein, Ph.D., Blacksburg, VA
Consulting Assistant
-developed and conducted performance appraisal/ merit pay training procedures for managers at Montgomery County Regional Hospital.
-as job analyst, collected data via interviews and questionnaires for the development of a performance appraisal and merit pay system.

8/90 - 12/90 Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, VA
Teaching Assistant for Research Methods graduate-level course.

4/90 - 5/90 Neil Hauenstein, Ph.D. & R.J. Harvey, Ph.D., Blacksburg, VA
Consulting Assistant
-as job analyst, collected data via questionnaires for the development of a performance appraisal and compensation system for Shenandoah Life Insurance Co.

1/90 - 5/90 Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, VA
Teaching Assistant and Laboratory Instructor for Quantitative Topics undergraduate-level course.

8/89 - 12/89 Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, VA
Teaching Assistant for Social Psychology undergraduate-level course.

5/89 - 8/90 Department of Entomology, Virginia Polytechnic Institute and State University, Blacksburg, VA
Database Assistant
-Monitored data quality and provided feedback reports to personnel for 25,000-record database.
-Implemented computer programs to transform data for spatial representations and mapping.
-Developed and conducted training procedures for data collection field personnel.

4/89 INMAR, Inc., Winston-Salem, NC
Research Assistant
-Assisted in the testing of employee job attitudes.

2/89 H. John Bernardin, Ph.D., Roanoke, VA
Consulting Assistant
-Designed preliminary methods for increasing utilization of employee healthcare system at Norfolk Southern Railroad.

8/88 - 5/89 Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, VA
Discussion Section Instructor for Introductory Psychology undergraduate-level course.

9/86 - 5/88 Department of Child and Family Psychology, Brown University Program in Medicine/ Rhode Island Hospital, Providence, RI
Research and Clinical Assistant
-Supervised research data scoring, coding, computer entry, and statistical analyses.
-Designed databases for research studies.
-Assisted psychologists in the administration of research procedures and clinical interventions.

PUBLICATIONS

Spirito, A., Stark, L.J., Grace, N., & Stamoulis, D. (in press), Common problems and coping strategies reported in childhood and early adolescence. Journal of Youth and Adolescence.

MANUSCRIPTS UNDER REVIEW

Harvey, R.J., Becker, R.L., Brill, R.T., Lawless, W., Murry, W.D., & Stamoulis, D.T. Dimensionality of the Myers-Briggs Type Indicator: Evidence from confirmatory and exploratory factor analysis. Manuscript submitted for publication to the Journal of Applied Psychology.

Hauenstein, N.M.A., Brill, R.T., & Stamoulis, D.T. The moderating effect of "halo error" on test validity: Real or artifact? Manuscript submitted for publication to the Journal of Applied Psychology.

PAPER PRESENTATIONS AT MEETINGS AND SYMPOSIA

Hauenstein, N.M.A., Brill, R.T., & Stamoulis, D.T. (1991, August). The moderating effect of "halo error" on test validity: Real or artifact? Paper will be presented at the meeting of the American Psychological Association, San Francisco, CA.

Brill, R.T., & Stamoulis, D.T. (1991, March). Applied behavioral analysis of interventions in the workplace targeting alcoholism and drug abuse. In T.D. Ludwig (Chair), Behavioral community psychology: Intervention and assessment programs to improve personal health and safety. Symposium will be conducted at the meeting of the Southeastern Psychological Association, New Orleans, LA.

Spirito, A., Stark, L.J., Grace, N., & Stamoulis, D. (1990, August). Common problems and coping strategies reported by children and adolescents. Paper presented at the meeting of the American Psychological Association, Boston, MA.

Stark, L.J., Spirito, A., & Stamoulis, D. (1988, November). Problems and coping of childhood. Paper presented at the meeting of the Association for the Advancement of Behavior Therapy, New York, NY.

Spirito, A., Stark, L.J., Williams, C.A., Stamoulis, D., & Axelson, D. (1988, March). Coping strategies utilized by referred and nonreferred pediatric patients and a healthy control group. Paper presented at the meeting of the Society of Behavioral Medicine, Boston, MA.

RELEVANT GRADUATE COURSES

Statistics (2 course sequence)
Multiple Regression
Multivariate Analyses
Advanced Topics in Psychometric Theory
Research Methods in Psychology
Quantitative Topics and Research Methods in Applied Psychology

Learning/ Information Processing
Social/ Personality Psychology
Organizational Theory/ Leadership
Work and Motivation
Personnel (2 course sequence)
Job Analysis and Job Evaluation
Contemporary Topics in Industrial/ Organizational Psychology

Dean Thomas Stamoulis