3-1988

Eliminating Sex Bias through Rater Cognitive Processes Training

Carter Ard
Western Kentucky University

Ard,

Carter H.

1988

ELIMINATING SEX BIAS THROUGH RATER COGNITIVE PROCESSES TRAINING

A Thesis

Presented to

the Faculty of the Department of Psychology

Western Kentucky University

Bowling Green, Kentucky

In Partial Fulfillment

of the Requirments for the Degree

Master of Arts

by

Carter H. Ard

March, 1988

AUTHORIZATION FOR USE OF THESIS

Permission is hereby

☑ granted to the Western Kentucky University Library to make, or allow to be made photocopies, microfilm or other copies of this thesis for appropriate research or scholarly purposes.

☐ reserved to the author for the making of any copies of this thesis except for brief sections for research or scholarly purposes.

Signed Carter H. Ard

Date 4/8/88

Please place an "X" in the appropriate box.

This form will be filed with the original of the thesis and will control future use of the thesis.

ELIMINATING SEX BIAS THROUGH RATER

COGNITIVE PROCESSES TRAINING

Recommended March 11, 1988
(Date)

Ray M. Mendel
Director of Thesis

Elizabeth S. Effinger

Harold L. Rowton

Approved May 5, 1988
(Date)

Elmer Gray
Dean of the Graduate College

Table of Contents

List of Tables

ELIMINATING SEX BIAS THROUGH RATER

COGNITIVE PROCESSES TRAINING

Carter H. Ard March, 1988 66 pages

Directed by: Ray M. Mendel, Elizabeth S. Erffmeyer, and

Daniel L. Roenker

Department of Psychology Western Kentucky University

The success of Rater Cognitive Processes Training as a strategy for eliminating sex bias in ratings of performance in a physically demanding job was investigated in the present study. One hundred undergraduate students from a mid-sized regional university served as subjects. The independent variables were type of training and sex of the ratee, resulting in a two by two factorial design. The dependent variable was the performance ratings assigned by the subjects. Subjects in the experimental condition were trained to recognize the important dimensions of performance for the job of feed handler and received one practice/feedback session. Subjects in the control condition completed a case study exercise in lieu of training. All subjects then viewed a videotape showing a feed handler moving and stacking what appeared to be 25 lb. bags, and afterward assigned ratings using a graphic rating scale. An ANOVA revealed a significant main effect for sex ($p < .026$) and a significant main effect for training ($p < .013$). The interaction between sex and training was not significant. Results indicated that Rater Cognitive Processes Training was

not effective in eliminating sex bias. Instead, a clear contrast effect emerged. Potential implications of this study and future research directions are subsequently explored.

Eliminating Sex Bias Through Rater Cognitive Processes Training

Introduction and Literature Review

Despite its subjectivity, performance appraisal, most often in the form of a rating, is the cornerstone of much personnel practice and research. For example, 89 percent of the police departments in major metropolitan areas use supervisory ratings as the primary method of performance appraisal (Landy & Farr, 1976). In addition, Landy and Trumbo (1980) determined that 72 percent of the validation studies published in the Journal of Applied Psychology between 1965 and 1975 used performance ratings as the criteria.

Because performance ratings are subjective, they are prone to systematic errors, such as halo, central tendency, and leniency (Smith, 1986). One particularly troubling problem is the presence of either intentional or inadvertent bias in performance ratings. There has been extensive documentation, for instance, of the fact that, given identical performance, raters have a general tendency to give men more favorable evaluations than women (Nieva & Guteck, 1980).

The fact that performance appraisal ratings can be biased against protected groups often leads to equal employment opportunity problems (Dipboye, 1985). The potential repercussions of an unfair performance appraisal

system has forced personnel researchers and practitioners alike to seek a solution to this problem. As a result, many techniques, ranging from altering the rating format to diary-keeping, have been tried to eliminate bias and inaccuracy in performance appraisal.

One of the most promising lines of research aimed at improving performance appraisal focuses on the cognitive processes of persons evaluating performance. The rationale behind this approach is that the best way to eliminate bias is to study its source--that is, before researchers can eliminate the prejudices of performance evaluators, they must understand how raters assess performance and at what point bias enters into the process. Once a rater's cognitive processes are understood, solutions to the problem of bias can be developed.

One solution that has grown out of the cognitive processes research is rater training. The focus here is to improve, or at least alter, the cognitive processes of persons evaluating performance so that the ratings they assign are accurate and fair. Rater training programs that incorporate cognitive processes have been largely successful in eliminating rater inaccuracy (Smith, 1986; Pulakos, 1984; Pulakos, 1986; McIntyre, Smith, & Hasset, 1984), but these programs have not been used for the purpose of removing bias from performance evaluations. Therefore, as a new test of the effectiveness of rater cognitive processes training,

a rater training program designed to eliminate bias will be tested and evaluated in the present study. Specifically, the researcher in this study will attempt to remove sex bias from performance ratings in a physically demanding job.

Bias, as defined here, refers to the systematic deviation of performance ratings from the true score. The terms "biased" and "innaccurate" are not interchangeable. Accuracy refers to the degree to which a rater's evaluation of a ratee approximates that ratee's objective (true) performance. When a subject rates performance innaccurately, his or her ratings will be randomly distributed around the true score. If, on the other hand, a subject's evaluation of a ratee's performance is biased, his or her ratings will be consistently skewed to one side of the true score. For example, if a rater has a negative bias concerning a woman's ability to do physically demanding work, he or she will consistently rate the woman lower than what her actual performance would justify.

Although accuracy is made up of several components (i.e., elevation, differential elevation, differential accuracy, and steretype accuracy), differential accuracy appears to be the most appropriate for assessing the accuracy of performance judgements (Borman, 1977). Differential accuracy measures the degree to which a rater is sensitive to ratee differences in patterns of performance across rating dimensions (Pulakos, 1986). Differential accuracy is

computed by correlating the ratings provided by subjects with the "true scores" provided by the expert raters.

Accuracy is the critical criterion for judging the quality of performance ratings (Borman, 1977). However, if one is to evaluate the accuracy of performance ratings, the true score of the ratee must be known. While it may be possible to determine a ratee's true score in the laboratory, it is impractical, if not impossible, to do so in a real organizational setting. One attractive alternative to assessing and improving rater accuracy in organizations would be to determine whether bias is present in ratings and to remove this bias through training. While this approach would not ensure that ratings were accurate, it would at least guarantee that males and females (and other protected groups) are not rated systematically too high or to low, thus warding off EEOC problems. The goal in the present study, therefore, is to demonstrate that through cognitive processes-oriented training, researchers can reduce or eliminate bias in performance ratings.

In the literature review that follows, the presence of sex bias in performance appraisal will be documented, and two major attempts to eliminate this problem will be explored. This section continues with the introduction of a model of performance appraisal which attempts to explain the cognitive processes of the rater. Rater training programs consistent with the model will subsequently be explored.

Finally, a new test of the effectiveness of rater cognitive processes training will be presented: that of eliminating sex bias from ratings of physically demanding task performance.

Sex Bias in Performance Appraisal

Pro-Male Evaluation Bias

Pro-male bias in performance appraisal is a well-documented problem. Bias against women has been found in ratings of the quality of their essays (Cline, Holmes, & Werner, 1977; Goldberg, 1968; Issacs, 1981; Toder, 1980), how well they relate to customers and other employees (Rosen & Jerdee, 1974), and their contributions to a group discussion (Taylor & Falcone, 1982). Furthermore, male applicants, in a study done by Guteck and Stevens (1979), received more positive ratings than female applicants in terms of acceptability, service potential, and longevity. The problem of pro-male evaluation bias is not limited to performance ratings, however. Terborg and Ilgen (1975), for example, found that while male and female applicants were rated as equally suitable for an engineering position, males were offered higher starting salaries and were assigned to more challenging positions than were female applicants.

Contrasting Findings

Despite the evidence of bias against women, a substantial number of investigators have found no differences in the performance evaluations of men and women (Frank & Drucker, 1977; Hall & Hall, 1976; Heilman & Guzzo, 1978;

London & Stumpf, 1983; Penley & Hawkins, 1980; Rose & Stone, 1978; Stumpf & London, 1981; Isaacs, 1981). Hall and Hall (1976), for example, found that when subjects read an identical description of either a male or female personnel director handling the problem of a vacancy in the production department, subjects rated the female director no differently than the male. Stumpf and London (1981) had subjects rate a candidate's suitability for promotion to fill a managerial vacancy. Here again, ficticious male and female candidates were perceived as equally qualified to fill the vacancy. Using an in-basket technique, Frank and Drucker (1977) found that subjects rated males and females no differently in terms of communication, sensitivity, planning, and organization.

Pro-Female Evaluation Bias

To further complicate the picture, several researchers found evidence of a pro-female evaluation bias. Abramson, Goldberg, Greenberg, and Abramson (1977), for example, found that female attorneys and paralegal workers were rated as having more vocational competance than identical males. Furthermore, Bigoness (1976) and Hamner, Kim, Baird and Bigoness (1974), found that females received higher ratings than did males in low-skilled and semiskilled jobs.

Sex-Role Congruence Explanation for Sex Bias

One reason for the mixed findings is due in part to the sex stereotype of the job in question. Rosen and Jerdee (1973, 1974a, 1974b, 1975), for example, have conducted a

number of experiments examining the influence of sex role congruence on performance ratings which confirm this hypothesis. The general conclusions obtained from these studies is that men will receive higher ratings than women on tasks congruent with expectations of appropriate mascline behavior while women will receive higher ratings on tasks congruent with expectations of appropriate feminine behavior. Numerous studies confirm Rosen and Jerdee's findings (Levinson, 1975; Cohen & Bunker, for example). When women or men violate a rater's stereotype they are much more likely to receive low ratings, despite the fact that their performance was really identical to their counterparts of the opposite sex who behaved in a sex role congruent manner. In those jobs that are less sex specific, such as college professor or personnel director, sujects tend to rate males and females the same.

Contrasting Findings

Not all research supports the sex role congruence hypothesis, however. Mai-Dalton, Feldman-Summers and Mitchell (1979), for example, have found that females who act "out of role" by behaving aggressively are evaluated more favorably than those who comply with conventional sex role stereotypes. Furthermore, Jacobsen and Effertz (1974) found that male leaders were evaluated more negatively than were female leaders, while male followers, on the other hand, received higher ratings than female followers. Obviously, a

simple gender role stereotype explanation does not account for all of the research results and more complex models of sex bias are needed.

Pro-Male Evaluation Bias in Physically Demanding Jobs

Despite the mixed findings regarding sex bias, there is one area where pro-male evaluation bias is both striking and persistent. Despite the fact that women have made significant strides in gaining entry to traditionally male white collar fields such as law, medicine, business, and engineering (Deaux, 1984), they remain frustrated in gaining entrance and advancement in those jobs that are considered to be physically demanding, such as the job of firefighter or police officer. While much more research has been conducted examining bias in professional fields, those researchers that have focused on blue collar jobs have indeed found the presence of much sex discrimination. For example, Harlan and O'Farrell (1982) found that in blue collar jobs not traditionally held by women but made available through affirmative action, female new hires are typically placed in jobs requiring much less skill regardless of the women's qualifications. Harlan and O'Farrell also found that women were also promoted at a much lower rate than males. Potts (1983) and Townsey (1982) have noted that women are especially under-represented in our nation's police force. In a more recent study, Hill (1987) found that in a job which required employes to lift, move and stack 25 pound bags of

feed, subjects rated the female lower than the male despite the fact that performance was identical.

Sex Difference in Physical Strength

Why is it that women are meeting such resistance in gaining entrance to these blue collar jobs? One reason is that there exists a real sex difference in physical strength between men and women. Men are, on the average, physically stronger than women (Astrand & Rodahl, 1977; Campion, 1983). Thus a person's stereotyped belief than in general men are more suitable and would perform better than women in these physically demanding jobs may be at least partially correct. Because this stereotype has been reinforced through observation and direct experience, bias that occurs in these cases would probably be much more difficult to eliminate. Therefore, women that could perform as well or better than most men are typically passed over during hiring or are given poor performance appraisals and lower salaries (Cassell, Director, & Doctors, 1975; Harlan & O'Farrell, 1982; Deaux, 1984; Deaux & Ullman, 1983).

Solutions to the Problem of Rater Bias in Evalution

Improving the Rating Scale

Affirmative action can be quite effective in eliminating sex bias in hiring practices, but the problem of bias in performance appraisals for physically strenuous jobs still remains. Many solutions have been tried to remedy the problem of bias in performance appraisal, with mixed success.

Please note, however, that researchers typicaly focused on eliminating such biases as central tendency, leniency, and other common rater errors, and not on eliminating sex or racial bias. Most recent attempts to eiminate the rater bias mentioned above involved developing the "perfect" rating format: one that would focus the attention of the raters on the important dimensions, thus eliminating systematic error. Unfortunately, little progress has been made despite 30 years of effort in this area (Landy & Farr, 1980). Research shows that even the most sophisticated rating scales (i.e., BARS, behavior summary scales, etc.) are no better than a carefully construced graphic rating scale in reducing common rater errors (Borman, 1979). Furthermore, changes in rating format have only slight impact on the accuracy of ratings (DeNisi, Cafferty, & Meglino, 1984).

Rater Error Training

Another technique designed to eliminate bias focused on training raters to recognize and avoid common rater errors, such as halo and leniency. These training sessions usually consisted of a lecture explaining various rater errors to the trainees, followed by group discussion, practice and feedback. While these programs were successful in eliminating some errors, such as halo and leniency (Latham, Wexley, & Pursell, 1975), other errors persisted, or even increased following training (Borman, 1978). In addition, the overall accuracy of the performance ratings did not

improve (Borman, 1975, 1979; Pulakos, 1984).

Cognitive Processes of the Rater in Performance Appraisal

A more promising line of research shifts the focus from the rating scale and rater errors to the persons actually evalating performance. Psychologists following this line of thought believe that before one can eliminate bias in performance appraisal, it is necessary that the cognitive proesses underlying performance appraisal be better understood. DeNisi, Caferty and Meglino (1984) have proposed a model of performance appraisal which attempts to explain the cognitive processes of a person observing and evaluating behavior. Components of the model include

1. Observation of the behavior by a rater.
2. Formation of some cognitive representation of the behavior by the rater.
3. Storage of this representation in memory.
4. Retrieval of the stored information needed for a formal evaluation.
5. Reconsideration and integration of the retrieved information with other items of information available.
6. Assignment of a formal evaluation to the ratee using a suitable rating instrument.

Rater training is one method researchers have used to address the cognitive processes involved in performance appraisal. Each of the above steps proposed by DeNisi, et.

al., (1984) suggest ways in which Rater Cognitve Processes Training (RCPT) could be used to reduce or eliminate sex bias.

Observation of Behavior

All performance appraisal begins with observation, whether it is observation of behavior, units produced, number of customer complaints, etc. While observing performance, the rater takes an active role in determining which aspects of the available information will receive attention (DeNisi et. al., 1984). According to DeNisi et al., how a rater searches for information will determine what behavior is observed. Some of the determinants of what is looked for include preconceived notions about the probable qualities of a good or bad performer and the nature of the rating instrument (DeNisi et. al., 1984). Preconceived notions are part of the schemata formed by a rater and provide a framework for seeking relevant information and interpreting the incoming stimuli (DeNisi et. al., 1984). For example, a rater who has catergorized a worker in terms of a "good worker" schema seeks information to confirm his or her preconceived notion about what constitutes good performance. RCPT can provide an alternate schema, or framework, for the raters to use when searching for performance information. In other words, training can improve the quality of the information gathered by the observer by teaching raters to actively seek truly relevant information, rather than relying

on their own ideas about what constitutes "good" or "bad" performance. Before training, for example, a rater may believe that a "good" worker in an assembly plant is one who is male with a high school education, who "gets along" with others and does what is expected. In training, the instructor can tell the supervisors that the organization considers "good" performers to be those who arrive on time, produce 100 units or more per hour and abides by all safety regulations. Sex, education, and/or popularity, the raters can be told, are irrelevant. While training raters to pay attention to relevant information may not totally eliminate bias in and of itself, it is an excellent first step.

The nature of the rating scale, as stated before, also influences what information is sought about performance. According to DeNisi et al. (1984), the rating scale seems to direct the attention of the rater, guiding him or her to look for certain dimensions of behavior and not others. Also, different rating scales require the rater to assume different roles. For example, a behavior observation scale calls for the rater to be more of an observer of behavior rather than an evaluator. RCPT can thus reduce bias by familiarizing the observer with the important dimensions on the rating scale, by helping the trainees understand what these dimensions mean, and by providing examples of good and bad behavior underlying these dimensions. Finally, RCPT can also teach the rater exactly what his or her role is (i.e., evaluator

and/or observer), and how best to fulfill that role.

Encoding of Performance Information

After information has been gathered through observation, it must be encoded and stored into memory for later retrieval. During the encoding process, raw information is taken in and organized. "Since encoding is a type of translation or interpretation, the way it is performed has major consequences for the ultimate use of that information" (DeNisi et al., 1984, p. 376). There is evidence that schemata help raters organize or encode information into memory. In fact, several researchers (e.g., Kuiper & Rogers, 1979; Lord, Foti & Phillips, 1980; Taylor & Crocker, 1981) have shown that raters may utilize schemata instead of available information about a ratee. Schemata can also influence the way that behavior is interpreted. For example, a rater may have the simple schema that Joe is a poor worker. That schema is used as a way of organizing incoming information about Joe and can become a source of bias in that behavior that is actually an example of good performance, such as Joe's going to training sessions to increase job skill might be seen by the supervisor as Joe merely looking for an excuse to keep from working. RCPT provides trainees with an alternate framework to use when assessing performance. Because this alternate, more accurate, schema is used by the raters during observation, relevant behavioral information will be searched for and, subsequently, encoded.

Retrieval of Performance Information

Schemata are also critical to the retrieval of performance information, because they influence what is actually recalled about a ratee. Research has shown that raters recall the target person as belonging to a category (i.e., "good" or "bad" performer) (Nathan & Alexander, 1985), and schemata determine the category in which the ratee is placed. Also, Snyder and Uranowitz (1978), for example, have provided evidence that schemata essentially distort behavior to make it consistent with general impressions held about a ratee. In addition, Cohen (1981) found that people may "remember" schemata-consistent behavior that they never saw. By providing an accurate schema through training, one can help ensure that raters recall truly relevant dimensions of behavior, because, up to this point, the information that has been observed and encoded is based on an accurate framework. For example, providing subjects evaluating performance in a physically demanding job with the accurate schema that a "good" performer is one who can lift a 25 pound weight without strain helps insure that the subjects look for and remember whether the ratee strained to lift the weight, and not focus attention, for example, on whether the ratee was male or female.

Information Integration and Rating

The final step in the model involves combining and integrating the information recalled and forming an

evaluation in the form of a rating. How the recalled information is weighed is important to the overall quality of the evaluation. Research has shown, for instance, that people tend to give negative information more weight than it really deserves (London & Poplawski, 1976; Wyer & Hinkle, 1976). A rater cognitive processes training program could reduce or eliminate rater bias by showing the raters those variables that should be ignored (i.e., given a zero weight) during the performance appraisal process, such as sex and/or race, and how the dimensions that are important should be weighed.

Training Programs Consistent with Cognitive Model

Rater Accuracy Training

Several training programs have been designed that address the cognitive processes outlined by DeNisi et al., (1984). Most of these programs are designed to improve rater accuracy in evaluating performance, and are called Rater Accuracy Training programs (RAT). There are typically two approaches taken in RAT: Performance Dimension Training (PDimT) and Performance Standard Training (PStandT) (Smith, 1986). Performance Dimension Training attempts to improve accuracy by familiarizing raters with the important dimensions of performance. Performance Standard Training provides raters with a framework for evaluating performance. A frame of reference is achieved by presenting samples of job performance to trainees along with the "true" ratings

assigned to the performance by trained experts. While these techniques have been used separately, they appear to be most effective when used in combination (Smith, 1986).

Reasons for the Effectiveness of Rater Accuracy Training

Rater Accuracy Training that incorporates these two methods of training is successful for several reasons. According to DeNisi, Cafferty and Meglino (1984), a successful training program teaches "...raters to use better search and integration strategies. Such a training program acknowledges that raters use schemata to collect, encode, and retrieve information and helps raters to either abandon their incorrect schemata or modify them to make them more accurate" (p. 385). The crucial first step for helping raters develop a more accurate schemata is to thoroughly familiarize the raters with the important dimensions of performance (PDimT). By defining these behaviors early in the rating process, raters are able to attend to them while observing performance, and are thus able to make independent judgements without relying on global impressions or stereotypes (Smith, 1986). PDimT provides direction for raters, enabling them to focus on truly relevant dimensions of performance, and thus improving their information-acquisition skills.

Lenny, Mitchell, and Browning (1983) provide evidence confirming this hypothesis. In their research, subjects were asked to evaluate a male's or female's performance on either an intellectual test or a piece of artistic work. The

subjects were given either vague evaluation criteria or very specific guidelines outlining exactly what qualities they were to look for. As predicted, when rules for performance evaluations were vague, the subjects tended to rely on their own stereotypes when rating performance, and rated the female's performance lower than the male's despite the fact that the intellectual tests or artistic pieces were really identical. On the other hand, when guidelines were clear, sex bias disappeared.

Information integration and rating assignment are perhaps the most complex steps in performance appraisal, and addressing these steps is most crucial to an effective RAT program. When raters are asked to combine observations of specific behaviors into a single composite judgement, they rely on their own standards of effective performance unless provided with an alternate framework. Presenting performance standards (PStandT) to raters prior to the ratings process allows them to establish an accurate frame of reference on which to base their evaluations (Smith, 1986). The majority of studies have done this by allowing raters to compare their own ratings of sample performance to expert ratings or "true scores." In several of the studies, training also included a detailed description of behavior rationales justifying the expert ratings (Athey, 1983; Mcintyre, Smith, & Hasset, 1984; Smith, 1984). All the studies that used PStandT training reported increases in rater accuracy (Athey, 1983; Fay &

Latham, 1982; McIntyre et al., 1984; Pulakos, 1984; Smith, 1984).

Examples of Successful Rater Accuracy Training Programs

Pulakos (1984, 1986) and McIntyre, Smith, and Hasset (1984) provide excellent examples of successful RAT programs which incorporate both PDimT and PStandT training. In Pulakos' study subjects were asked to evaluate the videotaped performance of a manager dealing with a problem subordinate. McIntyre et al's. research required the subjects to evaluate a college professor's videotaped lecture. When these programs are compared the following characteristics emerge. First of all, trainees are lectured on the importance of paying close attention to ratee behavior in terms of relevant performance dimensions. The rating scale is then distributed and its underlying dimensions are reviewed and discussed. Trainees are then asked to generate examples of behavior corresponding to each level of performance under each dimension. When the trainees have a firm grasp on the meaning of each dimension and corresponding levels of performance, they view a videotape of a person performing a task and assign performance ratings using the rating scale. Afterwards, the trainees discuss the ratings they assigned and the rationale behind them. The true scores for each performance dimension, as determined by a panel of expert raters, are then revealed and a rationale for each true score is given by pointing out specific behaviors to which the

experts attended when assessing the performance. More discusion follows and trainees' questions are answered. By following this format, which provides an accurate framework for the raters to use when evaluating performance, both studies were able to significantly improve the accuracy of the ratings that the subjects assigned.

Rationale for Present Study

In the following study rater cognitive processes training will be applied to a new setting, one in which the potential for bias is stronger than in the studies discussed above. As stated earlier, sex bias can be a particularly troubling problem when the job in question has traditionally been considered "man's work" and is physically demanding. Bias that is present under these circumstances can be extremely difficult to eliminate, because the rater's stereotypes have been reinforced through observation and direct experience. The following study, therefore, evaluates the effectiveness of rater cognitive processes training in eliminating rating bias for a physically strenuous job where real sex differences in performance traditionally exist. In Pulakos' (1984, 1986) and McIntyre et al.'s (1984) research, the subjects may have been successfullly trained because the jobs in question may have been perceived as less sex-specific than a physically demanding job. Therefore these studies may not represent an adequate assessment of the effectiveness of RAT in the present contest.

The same characteristics of rater cognitive processes training that made RAT successful in improving accuracy should be effective in reducing or eliminating bias. The basic training content does not change; what does change is the goal of the training program (i.e., eliminating bias as opposed to improving accuracy). In the present RCPT program, subjects are familiarized with the important dimensions of performance (PDimT), just as they would be in RAT. PStandT, as stated before, focuses the rater's attention on the truly relevant dimensions of performance, thus substituting the rater's own stereotypes concerning good performance with a more accurate (and therefore less biased) framework. In addition, this RCPT program also presents performance standards to raters (PStandT), as does RAT, by allowing the raters to compare their ratings to the "true scores." A more detailed description of the RCPT program used in the present study will be provided in the Method section.

This study is an extension of the research conducted by Hill (1987), who documented bias against women in performance appraisal for a material handling job. In her study, subjects viewed actors on videotape lifting, moving and stacking what appeared to be twenty-five pound bags of feed. Hill manipulated the sex of the actors, body size (below average vs. above average), time the ratings were collected (immediately after viewing the videotape vs. one week later) and knowledge of actual feed sack weight (i.e., subjects

acting as controls knew that the sacks actually weighed three pounds, while the subjects in t[illegible] exp[illegible]imental condition were led to believe that the sacks [illegible]. An analysis of variance revealed a significant main effect for sex ($p < .001$). No other main effects or interactions in Hill's analysis were significant.

The purpose of the present study is to determine whether the sex bias documented by Hill (1987), where the job in question is male-dominated and a real sex and body size difference in performance traditionally exists, can be successfully eliminated using RCPT. It is hypothesized that there will be an interaction between training and the sex of the actors in the videotape. Specifically, there will be no difference in the ratings assigned to the female and male by those subjects that receive training, while those subjects that do not receive training will assign significantly lower performance ratings to the female than they will to the male.

Method

Overview of Experimental Design

The purpose of this study is to determine if RCPT (Rater Cognitive Processes Training) is effective in eliminating sex bias from performance ratings. The two independent variables in this study are training (no training versus Rater Cognitive Processes Training) and the sex of the ratees (male feed handler versus female feed handler), resulting in a two by two factorial design. The dependent variable is the performance ratings assigned by the subjects.

Subjects

Participants in the study were 100 undergraduate psychology students at Western Kentucky University.

Stimulus Material

The performance of the confederate male and female feed handlers had to be identical so that any differences in the performance ratings could be attributed to sex bias. Videotaping actors performing as feed handlers allows the researcher to standardize performance so that hopefully the only discernable differences in the tapes are the sex of the handlers. Two such videotapes developed by Hill (1987) were adopted for the present study. In these videotapes, actors posing as job applicants were shown lifting, carrying and stacking what appeared to be 25 lb. bags of feed. In both of

these videotapes performance was identical and standardized. The actors in the videotape were of similar body [redacted] similar clothing (i.e., jeans, a plaid shirt, and athletic rubber-soled shoes), maintained a neutral facial expression, and performed at the same rate. A third videotape, developed by the experimenter for use in Rater Cognitive Processes Training, showed a male dressed just as the actors described above. The only difference was that this ratee performed the task less effectively. Each videotape lasted approximately five minutes.

Rating Scale

The rating form developed by Hill (1987) was also adopted for the present study. This rating instrument was a simple graphic scale consisting of 10 items assessing various aspects of performance: 1) the amount of strain the actor exhibited, 2) the amount of effort exerted, 3) the degree to which the actor appeared to struggle, 4) the level of fatigue of the actor, 5) the approximate number of hours the actor could work without a break, 6) the total number of breaks the actor would probably take during an eight hour shift, 7) the probable number of bags the actor could move in an hour, 8) the actor's care in handling the bags, 9) the actor's overall performance, and finally, 10) the applicant's overall suitability for the job of material handler.

True Scores

The true performance scores of the below-average feed

handler [illegible] that subjects rating this tape as a part of Rater Cognitive Processes Training could receive feedback as to the accuracy of their ratings. True scores were also needed to confirm that the actor in the training tape was actually portraying below-average performance as the experimenter intended. Graduate students from the industrial/organizational psychology program at Western Kentucky University served as the expert raters. These individuals qualified as experts because of their familiarity with the literature concerning accurate performance evaluation and sex bias. Prior to rating the training videotape, the experts were told why they were rating the tape, and the procedure that would be followed in establishing true scores.

The procedure is described below. The experimenter remained in the room throughout the session, acting as a consultant and a discussion facilitator. Before rating the actors' performance, the expert raters read over each item on the rating scale. These raters then discussed each item one by one, defining the underlining dimension and generating examples of how an actor might behave for each level of performance (good, poor, average, etc.) under each dimension. The experts then viewed the videotape used in training and assigned their ratings individually. After the performance ratings were complete, each expert revealed the rating that he or she had assigned for the first performance item. These

ratings were written on a blackboard and discussion of the ratings followed. During the discussion, each expert offered the rationale behind his or her rating, citing the behaviors observed which lead to the performance rating assigned. Debate continued until a consensus was reached, and a true score obtained. In this way true scores and a corresponding rationale were obtained for each item on the rating scale.

Procedure

A workshop approach similar to the one developed by Pulakos (1984, 1986) and by McIntyre and Smith (1983) was adopted for the present study. All groups except the control received identical training. The control groups received no training. All subjects reported to a classroom in groups of approximately 25. The experimenter read the same introduction to each group (see Appendix A for introductory script).

Experimental Condition Procedure

After the introduction, subjects in the experimental group were told that they would receive training in performance appraisal (see Appendix B for training script). The trainees were instructed not to rely on their own ideas about what "good" performance might be. Instead, they were to evaluate the applicants based upon the dimensions of the rating scale. The rating scale for use in rating performance was then distributed to the subjects. Subjects were encouraged to ask questions at any time and to participate in

the discussion. The trainer read the first item aloud and asked the subjects to list examples of how a ratee might behave under a lot of strain, average strain, and no strain at all. These responses were later checked so that those subjects who answered carlessly or not at all (indicating apathy or lack of attention) could be removed from the data analysis. After the subjects had time to prepare, the researcher asked how a ratee exhibiting average strain might behave. After hearing the responses of the trainees, the trainer asked how an average ratee might compare to the probable behavior of a ratee under a lot of strain or no strain at all. After listening to the subjects' responses, and giving feedback, the experimenter physically demonstrated how the ratee might behave for each level of performance. Each item on the rating scale which addressed the behavior of the ratee (i. e., items 1, 2, 3, 4, and 8) was discussed in this fashion.

Three of the other items required the raters to estimate something about the ratee: the length of time the ratee could work without a break, the number of breaks the ratee would probably need to take during an eight hour workday, and the number of bags the ratee could move in an hour. In order to establish a common frame of reference for the raters to use when evaluating performance under these three items, the trainer announced the number of breaks taken, bags moved, etc., for an average, below average, and above average

worker. Finally, for the last two items, which require the raters to determine the overall suitability of the ratee for the job of material handler, the raters were told to look over their ratings, weighing each item equally and to generate their rating based on whether the ratee was average for most of the items, or above average, etc.

Following this lecture/discussion period, the subjects practiced rating a feed handler using the rating scale. The trainees viewed a videotaped performance and assigned their ratings. Following the rating, the trainer called upon some subjects to disclose their ratings. These ratings were written on a blackboard. The trainer then revealed the true score for each item on the scale. A rationale for each of the true scores was given by pointing out specific behaviors that the experts attended to when rating the dimension. More discussion followed, and all questions were answered.

Control Condition Procedure

After the introduction, control condition subjects were given an exercise to complete which, the experimenter stated, was to "help get them thinking about performance appraisal and its importance". The true purpose of the exercise was to induce a fatigue effect similar to what the subjects in the experimental condition would develop after 30 minutes of training. In the exercise, subjects read a case study (see Appendix D) and gave their recommendations as to what they would do if they were the main character. The exercise

lasted approximately 30 minutes.

Performance Rating Procedure for Both Conditions

After either receiving training (experimental condition) or completing a case study exercise (control condition), all subjects viewed either the male or female feed handler videotape. Before evaluating the performance of the feed handler, all subjects completed a work history questionnaire, designed to interfere with the encoding of the information just obtained about the performance of the ratee. Encoding interference was included in the study because it more closely represents what happens in real organizations--that is, typically managers observe the performance of their subordinates, but are not able to rate their performance immediately, nor are they able to rate without distractions or competing information. Also, past research has shown that bias is more likely to ocur when encoding is hindered (Cooper, 1981). After completing the work history questionnaire, subjects rated the performance of the feed handler they had just observed, using the rating scale developed by Hill (1987).

Results

Hill (1987) originally intended for the rating scale used in the present study to measure three constructs: effort, fatigue and performance. However, Hill's analysis of the internal consistency of the 10 performance items on the rating scale indicated substantial homogeneity among the items (alpha = .79). Because of the unidimensionality of the item ratings, all 10 were combined into a composite in the present study, resulting in a single composite rating for each feed handler. The composite rating a feed handler could receive ranged from 1 for poor performance, to 5, for superior performance.

The means and standard deviations for the subjects' ratings of the male and female feed handler in both the training and control conditions are presented in Table 1.

Table 1. Means and Standard Deviations for Performance Ratings by Sex of Feed Handler and Rater Cognitive Processes Training

	Experimental (Training)		Control (No Training)	
	X	SD	X	SD
Male Feed Handler	3.812	0.525	3.560	0.468
Female Feed Handler	3.592	0.626	3.268	0.628

As one can see, subjects across conditions rated the male feed handler higher than the female feed handler. Also, subjects who received training rated both the male and the female feed handler higher than their counterparts in the control condition. To determine the significance of the differences in the means, a 2 (training) by 2 (sex of feed handler) Between Groups Analysis of Variance (ANOVA) was performaned on the data, revealing two significant main effects. Table 2 is a summary of the ANOVA results.

Table 2. Analysis of Variance for Performance Ratings by Sex of Feed Handler and RCPT

	df	MS	F	p
Main Effects	2	1.856	5.793	0.004
Sex	1	1.639	5.114	0.026
RCPT	1	2.073	6.471	0.013
2-way Interactions	1	0.032	0.101	0.751
Sex x RCPT	1	0.032	0.101	0.751
Explained	3	1.248	3.896	0.011
Residual	96	0.320		
Total	99	0.349		

As the statistics suggest, subjects across conditions rated the male feed handler significantly higher than the female feed handler, resulting in a main effect for sex $F(1,99) = 5.114$, $p < .026$. Also, subjects who received training rated both the male and the female feed handler signficantly higher than their counterparts in the control condition, resulting in a main effect for training $F(1,99) = 6.471$, $p < .013$. However, contrary to our hypothesis, the interaction between the sex of the feed handlers and RCPT was not significant.

Discussion

Clearly, the hypothesis that RCPT would be effective in eliminating sex bias from the performance ratings assigned to feed handlers was not supported. Instead of eliminating sex bias, the RCPT program produced a clear contrast effect. According to Cascio (1982), a contrast effect can occur when subjects evaluate more than one ratee at a time. In this situation, as raters evaluate one candidate they tend to use the other candidates as the standard. "Who they (the subjects) rate favorably, then, is partly determined by others against whom the candidate is compared" (Cascio, 1982, p. 200). In addition, Wexley, Yukl, Kovacs, and Sanders (1972) found that the magnitude of contrast effects is greatest with applicants of intermediate suitability (as was the case here), and could account for as much as 80% of the variance in ratings.

In the present experiment, a contrast effect probably emerged because subjects who received the training first viewed a below-average performer. The raters then used this below-average performer as a standard to evaluate a second (male or female) feed handler. Although the second feed handler was really an intermediate performer, he or she looked to be slightly better than average in contrast to the below-average feed handler.

Why was the training program unsuccessful, producing only a contrast effect instead of reducing bias? The program should have been effective because it incorporated observation, practice and feedback (all the elements of a successful RCPT workshop) into its design. In the following section, several possible explanations for the present program's lack of success will be explored. These explanations will focus primarily on the design and context differences between this study and the two successful RAT training programs upon which this study was based. This section will continue with a discussion of other factors that may account for the current study's results and will end with recommendations for training program modifications and future research directions.

Comparison and Contrast between this Program and RAT Programs

Type of Scale Used

Upon examining the two studies, (e.g. Pulakos, 1986; McIntyre et al., 1984) on which this training program was based, several differences between those studies and the current study emerge. For example, subjects in Pulakos' (1986) and McIntyre et. al.'s (1984) research used a BARS scale when assessing the ratees, while in the present study a simple graphic rating scale was used. The use of a graphic scale should not be considered problematic however, because as mentioned in the literature review, past research has

shown that a BARS is no better in reducing rater error than any other type of rating scale (Landy & Farr, 1980; Borman, 1979; Bernardin, Alvares & Cranny, 1976; DeCotiis, 1977). In fact, the BARS scale has even been outperformed by the graphic rating scale in such areas as reducing leniency and improving discrimination among raters (Borman & Vallon, 1974). But despite the fact that research has shown that the BARS scale is no better than a graphic rating scale, RCPT was originally designed to be used in conjunction with a BARS scale. It is possible that a graphic rating scale is not well suited to RCPT. Future research should explore whether the success of RCPT is influenced by type of scale used.

Length and Intensity of the Training Programs

One other difference is the length and intensity of the training programs. In the Pulaskos' (1986) study subjects had two practice/feedback sessions in which to learn to recognize dimensions of performance. In addition, the training lasted approximately an hour and a half. In the present study, on the other hand, subjects had only one opportunity to practice and to receive feedback as to the accuracy of their ratings, in a training session lasting only thirty minutes. It is possible that one practice/feedback session did not allow the subjects ample opportunity to learn to rate a feed handler's performance without bias.

McIntyre et. al.'s (1984) RAT program, however, was successful in improving rater accuracy despite the fact that

subjects received only one practice/feedback session, lasting approximately 30 minutes. In addition, McIntyre et al. did not allow any discussion after the true scores were revealed to the subjects. In this case, one short session proved sufficient, as accuracy was significantly improved ($p < .05$). Therefore, the fact that subjects in the present study received a minimal amount of training may not be a major cause of its ineffectiveness. On the other hand, eliminating bias may be more difficult than improving accuracy, and thus more practice/feedback sessions may be necessary in order for training to be a success.

Nature of the Performance Task and Goal of the Research

Clearly, the major difference between these programs is the nature of the behaviors that the subjects were to rate and the goals of the research. In Pulakos' (1986) research subjects were asked to evaluate a manager's handling of a problem subordinate. In McIntyre et al.'s (1984) research, the raters assessed the quality of a professor's lecture. Neither of the jobs in the above studies is typically considered strongly sex-stereotypic, nor particularly prone to sex bias (Landy & Farr, 1978). In addition both of these studies were focused on improving the accuracy of performance ratings, and not on reducing bias. Specifically, Pulakos (1986) examined each of the four components of accuracy mentioned above. McIntyre et. al. (1984) focused specifically on differential accuracy and distance accuracy,

which measures the absolute value of the deviation of the obtained ratings from the true scores.

The present study, on the other hand, focused on reducing bias in the ratings of a task which was both male-dominated and in which sex bias had been previously documented. Had Pulakos (1986) and McIntyre et. al. (1984) centered their training program around a male or female-dominated job (whether it was physically demanding or not), the differential accuracy of the performance ratings may not have significantly improved after training. In another possible scenario, accuracy in performance ratings may have improved significantly after training, yet undetected sex bias may have continued to be present.

The Possible Permanence of Sex Bias in Physically-Demanding Jobs

It is possible that the researcher's failure to find a reduction in bias after training is due to the fact that sex bias in physically demanding jobs is so well instilled in subjects that almost any training program would be unsuccessful. Walter Lipman (1922), who coined the term "stereotype," observed "There is nothing so obdurate to education or to criticism as the stereotype" (p. 73). In noting the resistance of schemata to change, Taylor and Crocker (1981) suggest that one possible explanation is the failure of the rater to encode information that is incompatable with his or her schemata.

In support of this hypothesis, Mount and Thompson (1987) found that ratings are more accurate when the behaviors of the ratee are consistent with the expectations of the rater. One possible explanation is that behaviors that are expected have greater salience, and, as a result, are noticed and recalled more easily than unexpected behaviors (Zadney & Gerard, 1974). Leniency bias may be explained in an analoguous way. According to Mount and Thompson (1987), recall of ratee behavior may be influenced positively or negatively, depending on whether the ratee confirms a rater's expectation. When the behavior of the ratee is perceived to be consistent with rater expectations the category takes on a positive connotation because the behavior is viewed as acceptable or expected. The fact that the male feed handler was rated more leniently (or conversely, that the female feed handler was rated more severely) could be explained using a schemata-conguent model of performance appraisal.

In the most extreme case, subjects could actually be perceiving that the female feedhandler is performing as well as the male, yet still be unwilling to assign them equal ratings, asserting perhaps, that "women just shouldn't do that type of work." Given that the present study was conducted in a laboratory using a more sophisticated subject pool, this explanation, although possible, does not seem likely. Such an incident would be more likely to occur in a real organizational setting where performance ratings would

have a direct impact on salary increases and hiring/firing decisions. No amount of training would be effective if raters are resistant to the change.

Future Research Directions

More research is needed to confirm or discount the hypotheses presented above. The failure of the present training program could simply be a matter of insufficient training. Perhaps a more intensive training program, one with more practice/training sessions, is necessary for subjects to learn the differences in the behavior of poor, average, and excellent feed handlers. After having learned these differences, the subjects would then recognize that both the male and the female feed handler were both average performers and would have rated them the same.

Assuming that more intensive programs are unsuccessful, the question remains as to whether sex bias (or racial bias, age bias, etc.) is too ingrained to erase with a short term training effort. One interesting research direction would be to determine if RCPT would be successful in eliminating bias in a job where a real sex difference in performance does not exist. It is possible that subjects could be more easily trained when assessing performance in a job where they realize there is no physical limitations which would prevent females from performing as well as males. Lenny, Mitchell and Browning (1983) provide one example of reducing bias through training for a task where there are no real sex

difference in performance. It is possible that bias is difficult to reduce only when the subjects know by observation and direct experience that a female, on the average, could not perform the task as well as a male, as is the case in a truly physically demanding job.

Another research question that should be further explored is whether accuracy improves after RCPT in a job where bias has been documented. It is possible that RCPT improves the accuracy of performance ratings, while at the same time remaining ineffective in removing sex bias. If evidence such as this is found, researchers should probably look elsewhere for solutions to the problem of bias in performance appraisal.

Finally, the cognitive processes model proposed by DeNisi, Caferty, and Meglino (1984), while providing an excellent framework, needs further exploration so that each aspect of cognitive process underlying performance appraisal can be better understood. As researchers reach a better understanding of each component of the cognitive process, steps can be taken to further improve the training that raters receive. For example, the schemata-congruent model of performance appraisal discussed by Mount and Thompson (1987) could provide essential information about how schemata operate and how to alter a rater's biased view of the world.

Since ratings are often the only means available for establishing criterion performance scores against which to

validate selection, promotion, or other selection decisions (Borman, 1979), it is essential that research into improving the rating process continue. By examining the cognitive processes used by raters, significant progress may continue to be made in improving the quality of performance ratings. Important advances must be made, otherwise selection tests, particularly those for physically-demanding jobs, will be continually questioned by managers, employees, and the courts (Campion, 1983).

Appendix A

Introduction and Instructions for Experimental and Control Conditions

I'm conducting a study of the relationship between one's prior work history and performance ratings. Most performance appraisals are conducted by supervisors who rate their employees' performance. This is the type of appraisal we will be concerned with today.

What I'd like for you to do is to assume that you are Personnel Manager for Pan American Feeds, a large cattle feed supplier. You have an opening for the position of feed handler. The feed handler's most important job duty requires that the employee be able to, safely and without excessive strain, move feed bags over the course of an eight hour work day. Accordingly, a work sample selection test has been developed to help assess this ability. The test requires that the applicant move material for 30 minutes without a break. Because you do not have enough time to review each applicant, as Personnel Manager you have asked me to prescreen the applicants and to make videotapes of candidates performing the 30 minute work sample selection test. The tape(s) you will observe show(s)only the first two minutes and the last three minutes of the 30 minute test because these segments provide the most important information about

the applicant. These segments will allow you to compre how the applicant appears at the beginning of the session and at the conclusion of 30 minutes of continuous work. You will carefully view the tape, and afterwards you will be asked to rate the applicant on three characteristics: (a) the amount of effort exerted, (b) the degree of fatigue that is apparent, and (c) overall performance.

Before we begin and to help you get a feel for how physically demanding the job is, I'd like for each of you to come pick up or attempt to pick up just one of the bags you'll observe being lifted in the video. If you have back problems, you may not want to completely lift the bag. The important thing is that you get a feeling for how physically demanding their task is so I'd like for you to at least lift a corner of the bag. Pick the bag up and set it down. Be sure to set the bag down rather than dropping it because it's likely to burst if it's thrown around. As you lift the bag keep your back straight and bend only at the knee like this .... (Demonstrate proper lifting technique. Wait for each participant, in an organizaed fashion, to lift the bag of feed).

Appendix B

racy Training Procedure

Before you [illegible] the applicant, I am going to train you to rate performance as accurately as possible. As part of the training, I will review the rating scale, we will discuss it as a group, and then you will be allowed to practice by actually rating an applicant.

The important thing to remember here is that you must observe the performance of the applicant through the eyes of the organization you represent. In order to do this you must abandon your own ideas about what a good or poor performer may be like, and must concentrate instead on what your organization believes is important. Pan American Feeds has developed a rating scale which reflects what they believe is important to good performance as a feed handler. You are asked to evaluate performance based on this scale. (Distribute scale) Please feel free to ask questions at any time and to participate in the group discussion.

(Ask the subjects to write their names on the scale, then read aloud the directions.) The first four items here are really assessing similar things, that is effort and fatigue, so we will consider them together. (Read items.) The key to getting these items right is to focus on the actual behavior of the applicant. With that in mind I'd like

for each of you to write examples of how an applicant might behave for each level of effort on items 1 - 4. After a few minutes of this I will call on a few of you to give me your answers. Any questions? ..... How do you think a person under a lot of strain and exerting a lot of effort might behave?

Below Average - Hesitation before lifting
slow, shuffling walk
approximately 15 sec. carry and return time
out of breath
lifts the bag very slowly
may groan when lifting the bag
may drag the bag up his or her body
drops the bag quickly and straight down
leans back excessively
excessively supports the bag against the hips and/or legs while carrying
must squeeze bags tightly in order to hold on
holds arms rigidly while carrying the bags

Above Average - move the bags quite rapidly
approximate carry and return time of 6 sec.
lifts the bags quickly and easily
Can support the bags solely with the arms
places the bags on the ground gently

is able to bend arms at the elbow while carrying the bags.

relaxed grip on the bags

Average - moves the bags at a moderate pace

approximate carry and return time of 10 sec.

leans back slightly

rests bags somewhat against hips and legs

moderate grip

(Instructor demonstrates each level of performance)

The next three items require you to make an estimate concerning the applicant. (Read items.) In order to get these items right I need to tell you what's typical for an average, below average and above average worker.

Below Average - probably could not work 2 hours

- would have to take frequent breaks during the course of the 8 hour day
- could only move about 60 bags during an hour period.

Average - could probably work 3-4 hours before requiring a break

- would sometimes need a break during the day
- could move approximately 180 bags during an

hour period

Above Average - could work more than 5 hours without a break

- would rarly or never need a break during the course of an 8 hour day
- could move approximately 300 bags during an hour period

The next item also requires you to look at the behavior of the applicant. What would be the correct way to lift, move and set down these bags?

- lifts the bag gently off the ground instead of jerking.
- hold the bags carefully with a relaxed grip
- sets the bags down gently; does not drop the bags

The last two items ask you to rate the overall suitability of the applicant for the job of material handler. (Read items). What I want you to do here is to look over those ratings dealing with effort and fatigue, and make your rating based on whether the applicant exerted an average amount of fatigue, etc.

Are there any questions? Now I'm going to let you practice rating an applicant using the rating scale. Do not assign any ratings until the videotape is over. After you have seen the tape I will call on a few of you and have you tell the class your ratings. Do not be embarrased if your ratings are wrong. I would be surprised if everyone got it right the first time around. After a few of you have disclosed your ratings, I will give you feedback concerning how close you were. Are there any questions? (Subjects then view the training tape, make their ratings, and engage in a practice/feedback session. The practice scales are then collected.)

You are now ready to rate an actual applicant for the position of feed handler. Pay close attention to the behavior of the ratee, and do not assign any ratings until I give you the signal.

Appendix C

Performance Rating Procedure for both the

Experimental and Control Conditions

After answering any questions, the experimentor begins the videotape. When the break in the tape occurs, the experimentor says, "You'll notice that the film has been cut here. We're now observing the last three minutes of the test. (Experimentor waits for tape to end). "Before rating this person's performance, I'd like for you to complete a Work History Questionnaire. Please write your name in the space provided in the upper right corner. The reason for having you write your name on the questionnaire is to correlate your responses across the forms you will complete today. What I'd like for you to do is to describe the most physcially demanding work you have ever done. If you have had more than one physically demanding job, describe the one job that you feel was the most physically demanding. Include any volunteer work you might have done, any housework or farm labor, and any military experience (e.g. high school ROTC). Do not include sports as physically demanding work. Think in terms of work, not play. As you finish, please remain seated, and don't communicate with others. Are there any questions?

(After everyone has complete the work history

questionnaire, collect the questionnaire while handing out the rating form). I'm passing out the rating form now. Write your name in the upper right corner. Now, rate the performance of the applicant in the videotape. (Instructions are read to the control group). Please be as accurate as possible when making these ratings. As you finish, please remain seated, and don't communicate with others. Are there any questions. (When subjects have complete the form, subjects are told that they will receive a full explanation of the purpose and results of the study at a later data to be announced).

Appendix D

Name.....

Rating Form

Read the entire form before making any ratings. Then go back and read each item carefully. Pay close attention to the verbal descriptions on each scale. Answer by palcing an "X" on the line closest to the answer which best reflects your opinion. Be as accurate as possible.

EFFORT

While performing the task, the applicant appeared to be under

---

a lot of strain	some strain	average strain	little strain	no strain

The amount of effort required of the applicant to complete the task appeared to be

---

very low effort	low effort	average effort	some effort	very high effort

To complete the task, the applicant seemed to struggle

---

a great deal	somewhat	average	a little	not at all

FATIGUE

After performing the task, the applicant appeared to be

---

not at all tired	a little tired	average	somewhat tired	very tired

Time and motion studies have shown that material handlers can work continuously for 2 hours before requiring a break. If necessary, this applicant would be able to continue working for ..... beyond the 2 hours before having to take a break.

could not work 2 hrs	2-3 hours	3-4 hours	4-5 hours	more than 5 hours

During the course of an 8 hour workday, employees take "breather" breaks i.e., they rest at their work station, chat with fellow employees, etc. How often in a 8 hour shift would this applicant need to take this king of break?

frequently	occasionally	sometimes	rarely	never

PERFORMANCE

In my opinion, this applicant should be able to move ..... bags of material in a 1 hour period of time.

300	240	180	120	60

The applicant handled the bags in such a way that the bags would not burst.

strongly agree	agree	neutral	disagree	strongly disagree

Would you recommend this applicant for the position of material handler?

strongly not recommend	not recommend	neutral	recommend	strongly recommend

Overall, the applicant's performance was

superior	above average	average	below average	poor

Appendix E

Case Study

Name.....

Franklin Community College employed a staff of 40 teachers. It was a new college offering a two-year associate of arts degree in a variety of areas. The teachers reported to Louise Medwick, who was in charge of faculty personnel. Economic conditions at the college were not good. The college had to fight for its yearly budget from the state education association, and lately, education had not been a high priority item. The college had been told that due to cutbacks, 20 percent of the teachers must be laid off.

Part of Medwick's job was to conduct an annual performance appraisal of the teachers. She did not like this part of her job, but she knew it was critical. Her evaluations would be the main basis for the layoffs. Her boss, college president Fred Schweiker, was adamant about keeping the "best" faculty, and it was her job to determine who was best. There was also the usual concern over raises, as part of a teacher's raise was based on merit. This year, though, the stakes were a lot higher. Its one thing to get a 6 percent raise when you thought you deserved 8 percent; its quite another to get laid off. Medwick knew her decisions would directly and intimately affect the lives of eight

teachers. She personally knew and liked the teaching staff, which didn't help matters either. The ax was going to fall, and it was just a case of whose heads were going to roll.

Medwick also faced a somewhat peculiar situation that made matters more easy and more difficult at the same time. The faculty at the college was not unionized. Thus there was no formal labor contract covering layoffs. Some organizations used seniority as the basis for layoffs - the last person hired was the first laid off. While the college was not compelled to consider seniority in making layoffs, they could do so if they wished. The problem was Schweilker didn't want to consider seniority - he wanted those laid off to be the poorest performers, not just the newest staff members.

The other oddity was that because the college was less than three years old, none of the staff had tenure. Tenure could preclude the dismissal of those teachers who had it, but no one did. Medwick saw the situation as a curse and a blessing. Seniority and tenure couldn't be used to reduce the pool of teachers who could be laid off, and this made her task more difficult. At the same time, poorly performing teachers couln't hide behind seniority and tenure as reasons for their retention. Thus, everyone was thrown into the same pot. It was her job to give them all a fair shake.

Medwick knew all about the usual methods of appraising teacher performance, but she was very aware of the

limitations when so much was on the line. She had used student ratings in the past. However, many teachers felt they were little more than a popularity contest. At least that's what the teachers who got low ratings said. She also used peer ratings, but only to help teachers improve, not for administrative decisions. Just about everyone taught the same number of classes, so there was no point in simply counting classroom hours. Besides, it would be hard to convice Schweilker that the best teachers also taught the most classes. Last year she wanted to start a behavioral measure of teacher performance - critical incidents, rating scales, the whole bit - but the idea got scratched because of time and financial problems. She wished she had forced the issue, but now it was too late.

What ever method she used, she would have to be able to explain and defend it. She also knew she would take a lot of heat from those who got laid off. While Medwick accepted her task as a part of the responsibility that comes with the job, she wished she had more solid information to go on. Picking the best from the rest was complicated and she wasn't totally sure in specific terms what "best" was. Best lecturer, best grader, best advisor? Medwick also knew that while some appraisals simply got filed away, this one wouldn't. The lives of 40 teachers and their families were riding on her decision.

Question: If you were Louise Medwick, how would you assess the performance of the teachers?

Appendix F

Name.....

Work History Questionnaire

Describe the most physically demanding work you have ever done. Include volunteer work and military service (e.g. ROTC), but do not include sports. Read each question carefully before responding. Use the back of this form if you need additional space.

1. What type of work did you perform? (e.g. fast food restaurant, baby sitting, manufacturing, farm, etc.)

2. How long were you employed? (Give month & year)
   From: To:

3. Did you work full-time or part-time?

4. One the average, how many hours per week did you work?

5. Did the job require you to lift (pick up, move to another area and put down) or move (push or pull to another area) heavy objects?

   If so, how heavy, in pound, were to objects you lifted?

6. Did you lift/move objects continuously i.e., nonstop?

7. Did you lift the objects over your head?

8. How many feey did you move the objects?

9. How long were you required to work before you could take a rest break?

References

Abramson, P. R., Goldberg, P. A., Greenburg, J. H., & Abramson, L. M. (1977). The talking platapus phenomenon: Competency ratings as a function of sex and professional status. Psychology of Women Quarterly, 2, 114-124.

Astrand, P. O. & Rodahl, K. (1977). Textbook of work physiology (2nd ed.). New York: McGraw-Hill.

Athey, T. R. (1983). The effect of group size on rater training and rating accuracy. Unpublished manuscript, Colorado State University, Fort Collins.

Bernardin, J. (1979). Rater training: A critique and reconceptualization. Proceedings of the 39th Annual Meeting of the Academy of Management.

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

Bigoness, W. J. (1976). Effect of applicant's sex, race, and performance on employer's performance ratings: Some additional findings. Journal of Applied Psychology, 61, 80-84.

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

Borman, W. C. & Vallon, W. R. (1974). A view of what can happen when behavioral expectation scales are developed in one setting and used in another.

Campion, M. A. (1983). Personnel selection for physically demanding jobs: Review and recommendations. Personnel Psychology, 36, 527-550.

Cascio, W. F. (1982). Applied psychology in personnel management. Reston, VA: Reston Publishing Co.

Cash, T. F., Gillen, B., & Burns, D. S. (1977). Sexism and "beautiyism" in personnel consultant decision making. Journal of Applied Psychology, 62, 301-310.

Cassell, F. H., Director, S. M., & Doctors, S. I. (1975). Discrimination within internal labor markets. Industrial Relations, 14, 337-344.

Cline, M. E., Holmes, D.S., & Werner, J. C. (1977). Evaluations of the work of men and women as a function of the sex of the judge and type of work. Journal of Applied Social Psychology, 7, 89-93.

Cohen, C. E. (1981). Person categories and social perception: Testing som boundaries of the processing effects of prior knowledge. Journal of Personality

and Social Psychology, 40, 441-452.

Cohen, S. L., & Bunker, K. A. (1975). Subtle effects of sex-role stereotypes on recruiters' hiring decisions. Journal of Applied Psychology, 60, 566-572.

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

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

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.

Deaux, K. (1984). Blue-collar barriers. American Behavioral Scientist, 27, 287-300.

Deaux, K. & Ullman, J. C. (1983). Women of steel: Female blue-collar workers in the basic steel industry. New York: Praeger.

Dipboye, R. L. (1985). Some neglected variables in research on discrimination in appraisals. Academy of Management Review, 10, 116-127.

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

Frank, F. D. & Drucker, J. (1977). The influence of

evaluatee's sex on evaluations of a response on a managerial selection instrument. Sex Roles, 3, 59-64.

Goldberg, P. A. (1968). Are women prejudiced against women? Transaction, 5, 28-30.

Guteck, B. A. & Stevens, D. A. (1979). Differential responses of males and females to work situations which evoke sex-role stereotypes. Journal of Vocational Behavior, 14, 23-32.

Hall, F. S. & Hall, D. T. (1976). Effects of job incumbents' race and sex on evaluations of managerial performance. Academy of Management Journal, 19, 476-481.

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work sampling task. Journal of Applied Psychology, 59, 705-711.

Harlan, S. L. & O'Farrell, B. (1982). After the pioneers: Prospects for women in nontraditional blue collar jobs. Work and Occupations, 9, 363-386.

Heilman, M. E. & Guzzo, R. A. (1978). The perceived cause of work success as a mediator of sex discrimination in organizations. Organizational Behavior and Human Performance, 21, 346-357.

Hill, C. (1987). Effects of sex and body type on ratings of physically demanding task performance. Unpublished Masters Thesis, Western Kentucky University, Bowling

Green.

Isaacs, M. B. (1981). Sex role stereotyping and the evaluation of the performance of women: Changing trends. Psychology of Women Quarterly, 6, 187-195.

Jacobsen, M. B. & Effertz, J. (1974). Sex roles and leadership perceptions of the leaders and the led. Organizational Behavior and Human Performance, 12, 383-396.

Kuiper, N. A. & Rogers, T. B. (1979). Encoding of personal information: Self-other differences. Journal of Personality and Social Psychology, 37, 499-514.

Landy, F. J. & Farr, J. L. (1976). Police performance appraisal. JSAS Catalog of Selected Documents in Psychology, 6, 83. (Ms. No. 1315).

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

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

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

Lenny, E., Mitchell, L., & Browning, C. (1983). The effect of clear evaluation criteria on sex bias in judgements of performance. Psychology of Women Quarterly, 7, 313-327.

Levinson, R. M. (1975). Sex discrimination and employment practices: An experiment with unconventional job enquiries. Social Problems, 22, 533-543.

Lipman, W. (1922). Public Opinion. New York: Harcourt, Brace.

London, M. & Poplawski, J. R. (1976). Effects of information on stereotype development in performance appraisal and interview contexts. Journal of Applied Psychology, 61, 199-205.

London, M. & Stumpf, S. A. (1983). Effects of candidate characteristics on management promotion decisions: An experimental study. Personnel Psychology, 36, 241-259.

Lord, R. G., Foti, R. J. & Phillips, J. S. (1980). A theory of leadership categorization. In J. Hunt and C. Schriesheim (Eds.), Southern Illinios University Leadership Symposium. Carbondale: Southern Illinois University Press.

Mai-Dalton, R. R., Feldman-Summers, S., & Mitchell, S. F. (1979). Effects of employee gender and behavioral style on evaluations of male and female banking executives. Journal of Applied Psychology, 64, 221-226

McIntyre, R., Smith, D. & Hasset, C. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied

Psychology, 69, 147-156.

Mount, M. K. & Thompson, D. E. (1987). Cognitive categorization and quality of performance ratings. Journal of Applied Psychology, 72, 240-246.

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

Nieva, V. F. & Guteck, B. A. (1980). Sex effects on evaluation. Academy of Management Review, 5, 267-276.

Penley, L. E. & Hawkins, B. L. (1980). Organizational communication, performance, and job satisfaction as a function of ethnicity and sex. Journal of Vocational Behavior, 16, 368-384.

Potts, L. W. (1983). Equal employment opportunity and female employment in police agencies. Journal of Criminal Justice, 10, 455-468.

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.

Rose, G. L. & Stone, T. H. (1978). Why good performance may (not) be rewarded: Sex factors and career development. Journal of Vocational Behavior, 12,

197-207.

Rosen, B. & Jerdee, T. H. (1973). The influence of sex-role stereotypes on evaluations of male and female supervisory behavior. Journal of Applied Psychology, 57, 44-48.

Rosen, B. & Jerdee, T. H. (1974a). Influence of sex-role stereotypes on personnel decisions. Journal of Applied Psychology, 59, 9-14.

Rosen, B. & Jerdee, T. H. (1974b). Effects of applicant's sex and difficulty of job on evaluations of candidates for managerial positions. Journal of Applied Psychology, 59, 511-512.

Rosen, B. & Jerdee, T. H. (1975). Effects of employee's sex and threatening versus pleading appeals on managerial evaluations of grievances. Journal of Applied Psychology, 60, 442-445.

Smith, D. E. (1986). Training programs for performance appraisals: A review. Academy of Management Review, 11(1), 22-40.

Snyder, M. & Uranowitz, S. (1978). Reconstructing the past: Some cognitive consequences of person perception. Journal of Personality and Social Psychology, 36, 941-950.

Stumpf, S. A. & London, M. (1981). Capturing rater policies in evaluating candidates for promotion. Academy of Management Journal, 24, 752-766.

Taylor, S. E. & Crocker, J. (1981). Schematic bases of social information processing. In E. Higgins, C. Herman, and M. Zanna (Eds.), Social cognition: The Ontario symposium (Vol. 1). Hillsdale, NJ: Erlbaum.

Taylor, S. E. & Falcone, H. T. (1982). Cognitive bases of stereotyping: The relationship between categorization and prejudice. Personality and Social Psychology Bulletin, 8, 426-433.

Terborg, J. R. & Ilgen, D. R. (1975). A theoretical approach to sex discrimination in traditionally masculine occupations. Organizational Behavior and Human Performance, 13, 352-376.

Toder, N. L. (1980). The effect of the sexual composition of a group on discrimination against women and sex-role attitudes. Psychology of Women Quarterly, 5, 292-310.

Townsey, R. D. (1982). Black women in american policing: An advancement display. Journal of Criminal Justice, 10, 455-468.

Tucker, D. H. & Rowe, P. M. (1977). Consulting the application form prior to the interview: An essential step in the selection process. Psychological Bulletin, 62, 283-287.

Wexley, K. N., Sanders, R. E., & Yukl, G. A. (1973). Training interviewers to eliminate contrast effects in employment interviews. Journal of Applied Psychology,

57, 233-236.

Wexley, K. N., Yukl, G. A., Kovacs, S. Z., & Sanders, R. E. Importance of contrast effects in employment interviews. Journal of Applied Psychology, 56, 45-48.

Wyer, R. S. & Hinkle, R. L. (1976). Informational factors underlying inferences about hypothetical people. Journal of Personality and Social Psychology, 34, 481-495.

Zadney, J. & Gerard, H. B. (1974). Attributed intentions and informational selectivity. Journal of Experimental Social Psychology, 10, 34-52.