5-1988

Effects of Gender & Body Size on Ratings of Physically Demanding Task Performance

Carolyn Hill
Western Kentucky University

Hill,

Carolyn A.

1988

EFFECTS OF GENDER AND BODY SIZE ON RATINGS OF PHYSICALLY DEMANDING TASK PERFORMANCE

A Thesis

Presented to

the Faculty of the Department of Psychology

Western Kentucky University

Bowling Green, Kentucky

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

by

Carolyn A. Hill

May 1988

EFFECTS OF GENDER AND BODY SIZE ON RATINGS OF PHYSICALLY DEMANDING TASK PERFORMANCE

Recommended 12/12/87
Date

Ray M. Mendel
Director of Thesis

Elizabeth S. Efflmeyer

Jamil L. Rankin

Approved 1-21-88
Date

Elmer Gray
Graduate College Dean

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to the following people without whom this paper would not be possible:

Dr. Ray Mendel, my advisor, for his advice, encouragement, counsel, and patience throughout my graduate career.

Dr. Betsy Erffmeyer for her advice and critical reading of the manuscript.

Dr. Dan Roenker for his assistance in data analysis and critical reading of the manuscript.

The actors who volunteered their time to assist in the making of the videotapes used in this research.

Nader Fotouhi for his support and encouragement throughout this project.

TABLE OF CONTENTS

LIST OF APPENDICES

TABLE

LIST OF TABLES

PAGE

LIST OF FIGURES

EFFECTS OF GENDER AND BODY SIZE ON RATINGS OF PHYSICALLY DEMANDING TASK PERFORMANCE

Carolyn A. Hill May 1988 62 pages

Directed by: Raymond M. Mendel, Elizabeth S. Erffmeyer, and Daniel L. Roenker

Department of Psychology Western Kentucky University

The purpose of this study was to examine the effects of gender and body size on ratings of physical performance and effort. Participants (N=250) viewed the videotaped performance of one of four actors (large man, small man, large woman, and small woman) lifting, moving, and stacking 25 pound bags of feed. However, instead of containing feed, the bags contained a light weight (three pound) packing material. Participants rated the actor's performance either immediately or one week after viewing the videotape. Although the actual performances were identical, a 2 (Gender) x 2 (Body Size) x 2 (Time of Rating) ANOVA revealed gender differences in performance ratings ($F(7,192) = 10.75$, $p < .001$). No differences were found between large and small individuals or between immediate and delay ratings. Implications of gender bias in performance ratings on physically demanding jobs are discussed.

Effects of Gender and Body Size on Ratings of Physically Demanding Task Performance

For the first time in its history, New York City employs women as firefighters. Their continued selection however is threatened by the potential validation of a selection test which places women at the bottom of the score distribution. The selection test is a maximum performance test. It consists of a series of work sample tasks that are scored in terms of the length of time required to complete the tasks. Jobs, on the other hand, rarely require maximum physical performance (Hogan, in press). Even most physically demanding jobs are performed at a submaximum level (Hogan).

One could argue that body size can be used to predict performance on a maximum physical performance test. In examining the relationship between height and weight and scores on physical fitness tests, Fleishman (1964) found a positive correlation. A logical extension of this finding is that it would be possible to predict a group of individuals' scores on a maximum physical performance test with some degree of accuracy based on body size. Given limited opportunity to observe performance, raters use whatever information is available when assigning ratings. One type of readily available information is body size.

Thus, raters may look at a small individual, assume his or her performance is lower than that of a large individual, and assign ratings accordingly. The same is true for large and small individuals within groups of males and females. In this way, within groups correlations are obtained--resulting in a spurious validity coefficient. Thus, large individuals, both male and female, score high on the selection test and receive high job performance ratings, whereas small individuals receive low test scores and low performance ratings. In this situation, there are two testable hypotheses: (a) the impact of body size on one's ability to judge test performance and (b) the impact of body size on the distortion of performance ratings.

The purpose of the present study is to address the second issue, namely the distortion of performance ratings. The goal is to examine the distorting effects of gender and body size on ratings of physically demanding task performance when true performance is known.

According to cognitive process theories (Feldman, 1981; Cooper, 1981; DeNisi, Cafferty, & Meglino, 1984) and the stereotype-fit model of discrimination (Dipboye, 1985), ratings reflect an objective aspect of the ratee's performance, but the ratee's performance is also evaluated on the basis of a stereotyped schema. Although several aspects of the ratee may influence ratings (e.g. likeability and attractiveness), it is likely that ratee gender and

body size are salient variables influencing ratings on a physically demanding task. The research on the cognitive processes involved in performance appraisals (Feldman; Cooper, 1981; DeNisi, Cafferty, & Meglino), the stereotype-fit model of discrimination in appraisals (Dipboye), and evaluation bias and attribution research will be employed to explore possible causes for the hypothesized discrepancies in ratings of male and female performance.

Most jobs that are considered physically demanding are found in traditionally male-dominated occupations such as construction, firefighting, and policing. Although these jobs are physically demanding to some extent, there is evidence that they are not as physically demanding as is commonly believed (Maher, 1984). For example, much of a police officer's time is spent in relatively sedentary activities such as riding in a patrol car and completing paperwork. A majority of the officer's time is not spent chasing criminals as television and movies lead one to believe (Maher). Furthermore, the job of homemaker, which is seen as "woman's" work and consequently not very physically demanding, has been shown to be as demanding as a police officer's job (Arvey & Begalla, 1975). However, the fact that physically demanding jobs are male-dominated leads to sex stereotypical attitudes about the jobs and the people required to fill them. The cognitive processes which underlie these stereotypes are similar to those involved in

job performance evaluations (Feldman, 1981). Therefore the literature regarding cognitive processes in performance appraisals is especially important in gaining an understanding of ratings of physical abilities.

Cognitive Processes in Performance Appraisals

Feldman (1981), Cooper (1981), and DeNisi, Cafferty, and Meglino (1984) propose similar but somewhat different models of the cognitive processes involved in performance appraisals. In summary, these models propose that four cognitive operations must be performed before performance appraisals can be made:

1. The rater must acquire information about the ratee through observing his or her performance.
2. This information must then be organized and stored for later access. New information must be integrated with old information.
3. The information must be retrieved from memory.
4. When a judgment is required, information must be integrated to form that judgment.

In the first step of the model, the rater acquires information regarding the ratee's job performance, but information unrelated to objective performance is also acquired. Ratee gender and body size are examples of such information. According to Feldman (1981) this information

is noticed automatically until it departs from expectations in which case controlled attention processes are employed. Thus, ratee gender and body size are detected automatically. However when a petite woman performs a physically demanding task, gender and body size are noticed through a conscious, controlled process.

According to DeNisi et al. (1984), the rater decides whether to attend to such information. One factor that determines the type of information to be noticed is the preconceived notions the rater has about the ratee. These preconceived notions affect what information is sought. For example, a rater who has categorized a small woman as weak will look for and notice the few occasions when she has to struggle to lift a heavy package but may not seek and attend to information indicating that she usually accomplishes the task with ease.

Whether ratee gender and body size are detected through an active cognitive process or a passive process is beyond the scope of the present study. The important fact is that these characteristics are retained and may be used when evaluations are made. According to cognitive process theories, the woman mentioned in the example above would receive a poor rating based on the rater's attention and retention of a few incidents which are unrepresentative of the ratee's typical behavior.

In step two, the information is encoded and stored for

retrieval at a later time. Information is not stored in its raw form (DeNisi et al., 1984). It is interpreted and stored in its interpreted or encoded form. Each incoming piece of information is assigned to a category. As the process continues, the rater begins to form a general impression about the ratee based on behavioral observations as tainted by preconceived notions or stereotypes, resulting in the ratee being categorized. When an evaluation must be made, this category is recalled, not actual behavior, thus leading to under and overevaluations of performance (Feldman, 1981). Therefore if a woman is categorized as weak, supporting evidence will be retained; and evidence which does not support the categorization will often be forgotten. If her "true" performance conflicts with the category, her rating will be an underevaluation of her performance.

During the encoding and storage phases, random and systematic decay occurs (Cooper, 1981). Systematic decay occurs in the direction of preconceived notions such that information inconsistent with these preconceived notions is more likely to decay than consistent information. The result is a stronger reliance on stereotypes when a performance evaluation is made. Therefore, the greater the time lapse between observation and evaluation, the more opportunity for information decay--both random and systematic, and the greater the influence of stereotypes on

ratings (Rosen & Jerdee, 1974a).

In summary, a major implication of the cognitive process models in the present context is that a rater will attend to a ratee's gender and body size to the extent the rater has gender and somatotypic stereotypes for physically demanding work, thus eliciting stereotypes regarding gender, body size, and physical abilities. As the rater observes the ratee's performance, new information about the ratee will be organized with the stereotype to form a category for that ratee. In addition, information decay will occur such that more stereotype consistent information will be retained than stereotype inconsistent information. Thus, a category for a ratee may be built around a stereotype of both women and small individuals as weaker than men and large individuals. When performance appraisals are made, objective information regarding the ratee is reconciled with the stereotype to form a summary judgment. The result is a lower rating for women and small individuals than for men and large individuals than may be justified by objective performance.

Stereotype-fit Model of Discrimination

Dipboye (1985) extends the hypotheses underlying the cognitive process models of performance appraisals to propose a model of discrimination in appraisals. Dipboye points out that discrimination is primarily a cognitive bias involving some of the same processes proposed by Feldman

(1981) and DeNisi et al. (1983) in their analyses of performance appraisals. According to Dipboye's stereotype-fit model, raters attribute to each ratee characteristics consistent with their stereotype of people who are similar to the ratee. For example, large individuals may be perceived as lacking in intellectual ability but endowed with plenty of physical ability. In addition, they attribute to a particular job, characteristics or requirements that are consistent with their stereotype of individuals who are successful at that job. For example, physically demanding jobs are considered "man's work" and are perceived as requiring masculine characteristics because it is assumed that men perform better than women at such jobs.

The essential point of the stereotype-fit model is that ratings reflect the rater's perceptions of the fit of the ratee to the stereotype of the job. Therefore, if a ratee does not fit the rater's stereotype of the job and of individuals in that job, he or she is more likely to receive an unfavorable rating. For example, a small individual does not fit the stereotype of a strong physical laborer. He or she would probably receive a poor performance rating because of his or her size rather than actual performance. His or her performance may be as good as or better than other ratees' but the stereotype distorts the perception of actual performance.

Darley and Gross (1983) examined the process leading to the confirmation of preconceived notions or stereotypes. Their study results revealed that raters actively search for evidence to confirm their stereotypes. Participants were indirectly given socioeconomic status (SES) information about a fourth grade child then shown a videotape of her performance on a standardized ability test. The performance was designed to give ambiguous information concerning the child's ability, i.e., she answered both easy and difficult questions correctly and incorrectly. Subjects who thought the child to be of low SES attended to information in the tape to confirm the stereotype that low SES individuals perform poorly on such tests, and those who thought the child to be of high SES attended to information to confirm stereotypes of standardized test performance of that group.

In a performance appraisal situation of a physically demanding job, a rater likely begins with a stereotype that men are stronger than women. He or she watches men and women perform a physically demanding task while searching for evidence to support his or her stereotype. This supporting evidence is retained and later recalled when evaluations are made. Because disconfirming information passes unnoticed or is forgotten, biased evaluations result.

Evaluation Bias and Causal Attributions

To this point, much has been said regarding differences between male and female performance and how those

differences are exaggerated through various cognitive processes. How does one explain differences in ratings of identical performance by males and females? What are the causal attributions made for such performances? The literature examining evaluation bias and casual attributions provides at least partial answers to these questions.

The experimental design used in most evaluation bias studies requires that participants read descriptions of hypothetical individuals who are identical except for gender. Participants are then asked to make evaluative judgments or personnel decisions regarding these individuals. In a review of this literature, Nieva and Gutek (1980) note that most of these studies have revealed pro-male evaluation bias. For example, Gutek and Stevens (1979) found that male applicants received more positive ratings than female applicants in terms of acceptability, service potential, and longevity. Schneier and Beusse (1980) found that managers in a performance appraisal training course rated female performance lower than male performance. Attempts to minimize bias by using a behaviorally-based format, while successful in many cases, were only partially successful here. Using an in-basket simulation rather than written descriptions, Terborg and Ilgen (1975) found that although male and female applicants were rated as equally desirable for an engineering position, the male was offered a higher starting salary. Participants

also assigned females to dull, routine jobs significantly more than to challenging, difficult jobs.

Rosen and Jerdee (1973, 1974a, 1974b, 1975) have conducted a number of experiments to examine the influence of sex role congruence on performance ratings. The uniting hypotheses in these studies is that women will receive higher ratings than men on tasks congruent with expectations of appropriate feminine behavior. Men are expected to receive higher ratings than women when the task is congruent with expectations of appropriate masculine behavior. For example, one study (1974a) found that a request for a leave of absence to care for small children was seen as significantly less appropriate for men than women.

Although Rosen and Jerdee's studies have consistently found that stimulus individuals receive higher ratings when the task is sex-role congruent, one study (1975) found that sex-role incongruent behavior was better received. When filing a grievance, aggressive, threatening behavior from a woman was better received, and a polite, pleading appeal was preferred from a man. This contradictory finding may be explained by the nature of the task i.e., filing a grievance. It may be that aggressive, threatening behavior from a woman provides information about the intensity of the complaint because it is inconsistent with commonly held sex-role stereotypes. The major implication from Rosen and Jerdee's research (1973, 1974a, 1974b, 1975) for the present

study is that women will receive lower ratings than men because the physical nature of the task is incongruent with sex role stereotypes of appropriate female behavior.

In contrast to the studies showing pro-male evaluations, a number of studies have found evidence of pro-female bias. In a field study, supervisors rated female performance higher than male performance (Mobley, 1982). A possible explanation given for this result was consequences to the rater for biased evaluations (e.g. employee signature on the rating, employee grievance procedure for perceived unfair ratings, and review of the ratings by upper management and internal EEO officers). Abramson, Goldberg, Greenberg, and Abramson (1977) found that both male and female participants rated a female attorney and paralegal worker as more competent than their identical male counterparts. They labeled this finding the "talking platapus phenomenon." The talking platapus phenomenon is manifest when an individual, especially a woman, achieves an unexpected level of success and as a result evaluations of her performance are magnified. "After all, it matters little what the platapus says, the wonder is that it can say anything at all." (1977, p. 123). The talking platapus phenomenon is not likely to arise in the present study for this reason the degree of success with which the ratee performs the task is not addressed; therefore the talking platapus phenomenon or pro-female bias is not expected.

The results from a number of studies have revealed no differences in male and female evaluations. Using an in-basket technique, Frank and Drucker (1977) found no differences in ratings of men and women on written communication, sensitivity, planning and organization. Hall and Hall (1976), using an extensive case study of a male or female personnel director, found no differences in ratings of motivation, ability, and overall task performance. A possible explanation for finding no gender differences in evaluations in this study is the amount of behavioral information provided to the raters. It appears that the more behavioral information available the less raters rely on stereotypes. Isaacs (1981) found no differences in male and female performance ratings in traditionally masculine fields once the woman had achieved status in that field.

Evaluations include not only judgments of the worth of the performance but also the causal attributions for the performance (Nieva & Gutek, 1980). Causal attributions of performance are important because they determine whether performance is seen as an accidental occurrence or consistent behavior. According to attribution theorists, performance can be attributed to four causes: (a) ability, (b) effort or motivation, (c) task difficulty, or (d) luck (Feldman-Summers & Kiesler, 1974). Ability and task difficulty are relatively stable factors, whereas effort or motivation and luck are unstable.

Most attribution research asks the participant to attribute performance to luck, effort, ability, and task difficulty by completing a continuous rating of each of the four factors (Deaux & Emswiler, 1974; Abramson et al., 1977; Hall & Hall, 1976; Isaacs, 1981). However, Feldman-Summers and Kiesler (1974) asked subjects to attribute identical male and female performance to a combination of these factors by dividing a circle into segments of the four attribution factors varying the size of each segment according to its relative influence. Etaugh and Brown (1975) asked participants to attribute performance to only one of the four factors. Regardless of the scale used, research indicates that stable factors are typically used to explain expected (male) success, and unstable factors are typically used to explain unexpected (female) success.

In a review of gender differences in attribution research, Ross and Fletcher (1985) note that successful male performance is more likely to be attributed to ability and less likely to be attributed to luck or effort than successful female performance. For example, Deaux and Emswiller (1974) found that successful performance by men was attributed to skill, whereas successful performance by women was attributed to luck. Attributions for failure are reversed (Ross & Fletcher). Male failure is more likely to be attributed to bad luck or lack of effort, and female failure is likely to be attributed to lack of ability. For

example, Etaugh and Brown (1975) found that failure by men, an unexpected outcome, was attributed to unstable factors such as a lack of effort or bad luck; and failure by women, an expected outcome, was attributed to lack of ability and task difficulty, stable factors.

In summary, the purpose of the present study is to examine the effects of gender and body size on physically demanding task performance ratings. Given equivalent objective performance, it is hypothesized that raters will rate the performance of men as less effortful on a physically demanding task than that of women. The underlying rationale is that observing the ratee and thereby noting his or her gender will stimulate the recall of common stereotypes regarding the relative physical abilities of the genders, i.e., that men are stronger than women (Fleishman, 1964). Raters will then attend to stereotype confirming information, disregard or forget nonconfirming information, and base ratings on a combination of gender bias and objective performance (DeNisi et al., 1984; Feldman, 1981; Cooper, 1981). Based on the same rationale, it is hypothesized that large individuals will receive higher performance ratings than small individuals. According to the cognitive process theories of performance appraisal, the greater the time delay between observation of objective performance and the assignment of ratings, the more raters base their ratings on stereotypes. Thus, it is hypothesized

that ratings obtained after a one week delay will reflect more gender and body size bias than those obtained immediately after viewing objective performance. The causal attributions for performance which are included in evaluations are important because they determine whether performance is perceived as consistent or an accidental occurrence. It is hypothesized that female performance will be attributed to unstable factors (luck or effort), whereas male performance will be attributed to stable factors (task difficulty or ability).

Method

Overview

A 2 x 2 x 2 between groups ANOVA was used to analyze the effects of gender, body size, and time of rating on ratings of physically demanding task performance. Four videotapes were prepared, each featuring one combination of large and small, male and female actors performing what appeared to be a physically demanding task. The bags being lifted in the videotapes weighed approximately three pounds, but participants in the experimental condition were led to believe that they weighed 25 pounds. Participants in the control condition were told the actual weight of the bags. After viewing one of the four tapes, participants rated the actor's performance either immediately upon completion of a Work History Questionnaire or after a one week delay.

Videotape Preparation

Videotapes were prepared featuring large and small, male and female actors performing what appeared to be a physically demanding materials handling task. Confederates were filmed lifting what appeared to be twenty 25 pound bags of feed, moving them a distance of 12 to 15 feet, and stacking them. However, instead of containing feed, the bags actually contained packing material and weighed approximately three pounds.

Four five-minute scenes were taped using a work sample selection test setting. The first two minutes of each scene consisted of 85 seconds of instructions from the experimenter followed by 35 seconds of the selection test. During the instruction segment, the experimenter, posed as a personnel assistant, explained the proper lifting technique and that the test required moving the bags for 30 minutes without a break. After 35 seconds during which the actor began the test by moving three bags, a segment of static lasting approximately one second was inserted to depict a "break" in the film. Participants were told that to save time they would view the first two minutes and the last three minutes of the test because these segments provided information about how the applicant appeared before and after 30 minutes of continuous work. Therefore, the "break" represented a 30 minute time lapse during which the actor continued to perform the test. The actor then moved 20 bags for the last three minutes of the tape.

Except for the gender and size manipulation, an attempt was made to make each of the four tapes identical by standardizing the setting and performance, especially the manner in which the bags were lifted and the time required to move them. Table 1 shows the mean and standard deviation in seconds for each of the four actors for: (a) the time required to carry a bag 12 to 15 feet (Carry), (b) the time required to return to pick up another bag (Return), (c)

Carry + Return (C + R), and (d) total moving time (Total). Carry and Return Times reflect the time required to walk 12 to 15 feet and do not include the time needed to pick up or put down a bag. Total moving time is the time required to move all 23 bags including walking, lifting, and setting down bags.

Table 1
Mean and standard deviation in seconds for each actor for carry, return, c + r, and total times.

	Carry		Return		C + R		Total
	X	SD	X	SD	X	SD	
Large Man	3.95	.61	2.34	.77	6.29	.89	215
Small Man	3.58	.44	2.43	.79	6.01	.85	214
Large Woman	3.97	.50	2.53	.83	6.46	.94	217
Small Woman	3.51	.61	2.43	.80	5.94	.87	207

As can be seen from Table 1, the differences across the four actors for Carry, Return, Carry + Return, and Total Times are extremely small. The maximum differences in mean Carry, Return, and Carry + Return Times are .46, .19, and .52 seconds, respectively. It is not likely that participants would be able to detect these extremely small differences. If participants did detect differences in Carry or Carry + Return Time, this would work against the hypotheses of small individuals and women receiving lower ratings because the small woman has the shortest Carry and Carry + Return Times. Although the small woman has an intermediate Return Time, the differences across the four

actors in Return Time is less than for either Carry or Carry + Return Time and would be the most difficult to detect. Although the 10 second difference across the four actors in Total Time is substantially larger than the differences in mean Carry, Return, and Carry + Return Times, it is still quite small and not likely to be detected over a three minute time span.

In addition to standardizing the time required to move the bags, particular care was taken to standardize the manner in which the bags were lifted. Standardization of the manner in which the bags were lifted was accomplished by instructing the actors in the proper lifting and carrying technique. The tapes were further standardized by using the same camera angle and distance from the actor for all four tapes. Moreover, the resolution of the tapes was such that facial expressions were not clearly defined, thus minimizing any contamination due to facial differences that may have occurred. The actors wore similar attire, i.e., jeans, plaid shirts, tennis shoes, and no heavy make-up or dangling jewelry.

Thus, the only substantive difference between the tapes was that the actors were varied to depict a large woman, a small woman, a small man, and a large man performing the task. Large and small body size were operationally defined using height and weight statistics obtained by the United States Department of Health, Education, and Welfare (1979)

(see Table 2).

Table 2
Weight, in pounds, and height, in inches, of adults aged 18 to 24 by gender. [1]

	Weight		Height	
	X	SD	X	SD
Men	165	29.6	69.7	2.8
Women	132	27.4	64.3	2.5

1 US Dept. of Health, Education, and Welfare (1979)

The four actors were between the ages of 18 and 24. A large man and large woman were selected so that their height and weight were approximately one standard deviation above the mean for their respective gender (see Table 3). A small man and small woman were selected so that their height and weight were approximately one standard deviation below the mean for their respective gender (see Table 3). One confederate from each of these four categories was videotaped

Table 3
Actors' weight, in pounds, and height, in inches.

	Weight	Height
Large Man	195	73
Small Man	135	67
Large Woman	165	71
Small Woman	104	62

To facilitate the perception of relative body size in

the videotapes, the experimenter and the actor appeared together during the instruction section of each tape. Participants saw the experimenter in person during data collection. In addition, the actor stated his or her height and weight and the experimenter repeated this information during the instruction section of the tape. Thus, participants were informed of the actor's height and weight, and they could compare the actor's body size to that of the experimenter.

Participants

Participants were 250 undergraduate psychology students at Western Kentucky University. Participation was voluntary.

Procedure

A 2 (Gender) x 2 (Body Size) x 2 (Time of Rating) Fractional Factorial design was used. Data were collected to complete all eight cells in the experimental condition. However, data for only the large man and small woman were collected in the control condition because the greatest differences in performance ratings were expected between the large man and the small woman. Finding no differences between these two groups in the control condition allows one to assume that no differences exist between any other groups in the control condition. Specifically, where subjects know the bags weighed only three pounds, and they reported no effort or performance differences between these most extreme

conditions (i.e., large man, small woman) it is reasonable to assume that no reported differences are likely between the remaining control group comparisons.

In the experimental condition, most of the data were collected during class time and treatment conditions were randomly assigned to classes. The experimenter read the same standardized instructions to each group (see Appendices A and B). Each group was shown one of the four videotapes under either a delay or immediate rating condition. The videotape and rating condition were selected at random, and the group was not informed of the existence of the other three tapes or alternate rating condition until debriefing.

Before viewing the videotape, the experimenter explained to the group that its task was to evaluate the performance of someone performing a materials handling task. The task was described as physically demanding; to illustrate this point, each participant lifted or attempted to lift a bag of feed similar to those shown on the videotape. However, here the bag actually contained feed and weighed 25 pounds rather than three pounds. After each participant lifted the bag of feed, the videotape was shown.

After viewing the videotape, the participants completed a Work History Questionnaire (see Appendix C). The purpose of completing the questionnaire was to interfere with the encoding, storage, and retrieval of the behavioral information presented on the videotape. Four of the eight

experimental groups completed the Rating Form immediately after viewing the videotape and completing the Work History Questionnaire. The other four groups watched the videotape and returned a week later to complete the Work History Questionnaire and the Rating Form. The purpose of varying the time between observation and rating was to test the hypothesis that raters rely more on stereotypes than objective performance when more immediate rating is not possible.

The Rating Form (see Appendix D) consisted of 10 items which were rated on a 5-point scale: 3 items assessed effort, 3 fatigue, and 4 performance. Upon completion of the Rating Form, participants were asked to complete a Supplemental Rating Form (see Appendix E) containing three items. On one item, participants attributed performance to one of four factors: (a) luck, (b) effort, (c) ability, and (d) task difficulty. The second item asked the participants to record the applicant's height and weight as a manipulation check on their height and weight perceptions. Similarly, the final item served as a check to determine whether the participants perceived the actors to be of above or below average body size.

The control conditions were administered after the experimental condition to prevent potential experimental group participants from learning the actual weight of the bags lifted in the videotapes.

The purpose of the control condition was to determine whether the ratings varied as a function of the perceived physical demands of the task. The experimental and control conditions were identical except that participants in the control condition were told that the bags in fact actually weighed approximately three pounds. Each participant in the control condition lifted a bag containing three pounds of packing material prior to viewing the videotape.

After all experimental data were collected, the participants were fully debriefed. The nature and purpose of the deception was explained, and all questions were answered.

To summarize, finding no differences in the ratings obtained across the four control groups while obtaining differences among the experimental groups would support the hypothesis that gender and body size influence the ratings of task performance only when the task is seen as physically demanding.

Analysis and Results

The first step in the analysis was to determine the dependent variable or variables. Although the Rating Form was originally designed to measure three constructs--effort, fatigue, and performance--the internal consistency was calculated, via Cronbach's alpha, to determine the dimensionality of the construct or constructs being assessed. Before calculating alpha for the Rating Form, items 2, 4, 7, 8, and 10 were reverse scored so that all items were rated on a 5-point scale ranging from (1) poor to (5) excellent. The item intercorrelations are shown in Table 4. Cronbach's alpha across all 10 items was .79 indicating a largely unidimensional set of items. Thus, a composite score based on the mean of all 10 Rating Form items was calculated and served as the dependent variable in all subsequent analyses.

Table 4
Item intercorrelations of rating form.

	Item									
Item	1	2	3	4	5	6	7	8	9	10
1	-									
2	.45	-								
3	.50	.45	-							
4	.52	.47	.54	-						
5	.28	.25	.26	.30	-					
6	.23	.29	.36	.33	.45	-				
7	.17	.15	.16	.19	.16	.20	-			
8	.30	.11	.22	.25	.14	.12	.18	-		
9	.31	.24	.26	.28	.15	.16	.22	.27	-	
10	.32	.22	.35	.30	.26	.18	.20	.28	.64	-

alpha = .79

The results of the manipulation check indicated that participants did detect a size difference among the four actors. Large individuals were rated as above average and small individuals rated as below average in height and weight compared to adult males and females. In addition, when asked to indicate the actor's actual height and weight baed on their recollection of the information given in the instruction section of the videotape, participants could do so with great accuracy (see Table 5).

Table 5
Actor's estimated height, in inches, and weight, in pounds, compared to actual height and weight.

Actor	Actual Height	Estimated Height	
		X	SD
Large Man	73	73.5	1.31
Small Man	67	68.1	1.89
Large Woman	71	69.6	1.98
Small Woman	62	62.9	2.00
Actor	**Actual Weight**	**Estimated Weight**	
		X	SD
Large Man	195	194.1	15.08
Small Man	135	141.3	11.68
Large Woman	165	160.1	9.22
Small Woman	104	110.8	11.57

The 2 (Gender) x 2 (Body Size) x 2 (Time of Rating) between groups analysis of variance (ANOVA) for the experimental data revealed a significant main effect for gender ($F(7,192) = 10.75$, $p < .001$ (see Table 6). The body size and time of rating main effects and the interactions were not significant.

Table 6
ANOVA summary table of composite rating by gender, body size, and time of rating for experimental data.

	df	MS	F	p
Main Effects				
Gender	1	2.69	10.75	.001*
Body Size	1	0.75	2.98	.086
Time of Rating	1	0.58	2.33	.128
2-way Interactions				
Gender x Size	1	0.00	0.00	.976
Gender x Time	1	0.42	1.69	.196
Size x Time	1	0.01	0.04	.842
3-way Interactions				
Gender x Size x Time	1	0.15	0.59	.445
Residuals	192	0.25		

*$p < .01$

Because the main effect for time of rating was not significant and because the greatest differences in ratings were expected between the large man and the small woman, only immediate ratings of the large man and the small woman were collected in the control condition. The large man/small woman differences were not statistically significant ($t = 1.34$, $p < .186$). Table 7 shows composite rating means and standard deviations for all experimental and control conditions and by gender and body size.

Table 7
Means and standard deviations of composite rating by gender and body size and for all experimental and control cells.

	Experimental		Control	
	X	SD	X	SD
Immediate	3.47	.48	3.70	.44
Men	3.54	.51	3.78	.47
Women	3.30	.49	3.62	.40
Large	3.54	.40	3.78	.47
Small	3.41	.55	3.62	.40
Large Man	3.64	.42	3.78	.47
Small Man	3.45	.51	3.62	.40
Large Woman	3.44	.35		
Small Woman	3.36	.59		
Delay	3.37	.54		
Men	3.53	.55		
Women	3.20	.49		
Large	3.42	.50		
Small	3.31	.60		
Large Man	3.56	.42		
Small Man	3.50	.67		
Large Woman	3.28	.53		
Small Woman	3.12	.44		

Although the primary purpose of the Work History Questionnaire was to interfere with the encoding, storing, and retrieval of information regarding the actor's objective performance, Pearson correlation coeficients were calculated to determine whether previous experience performing physically demanding work influenced the composite rating. The Pearson correlation analysis revealed no relationship between prior work experience and performance ratings.

The attribution data (item 1 of the Supplemental Rating Form) was analyzed using a chi square technique. No relationships were significant.

Discussion

The purpose of the present study was to examine the relationships between gender, body size, rating delay, and their interactions on ratings of physically demanding task performance. The hypothesis that women would receive lower ratings than men was confirmed. However, the second hypothesis that small individuals would receive lower ratings than large individuals was not confirmed. The hypothesis that ratings for women and small individuals would be lower as a function of the time delay between the observation of performance and the assignment of ratings was not confirmed.

In the present study, the hypothesis of a gender bias in physically demanding task performance ratings was supported. Female performance was rated lower than male performance (see Table 7). This finding supports the work of DeNisi et al. (1984), Cooper (1981), and Feldman (1981) on cognitive process theories in performance appraisals. Raters may well have attended to the ratee's gender, recalled the stereotype of men as physically stronger than women, and allowed these stereotypes to influence their ratings.

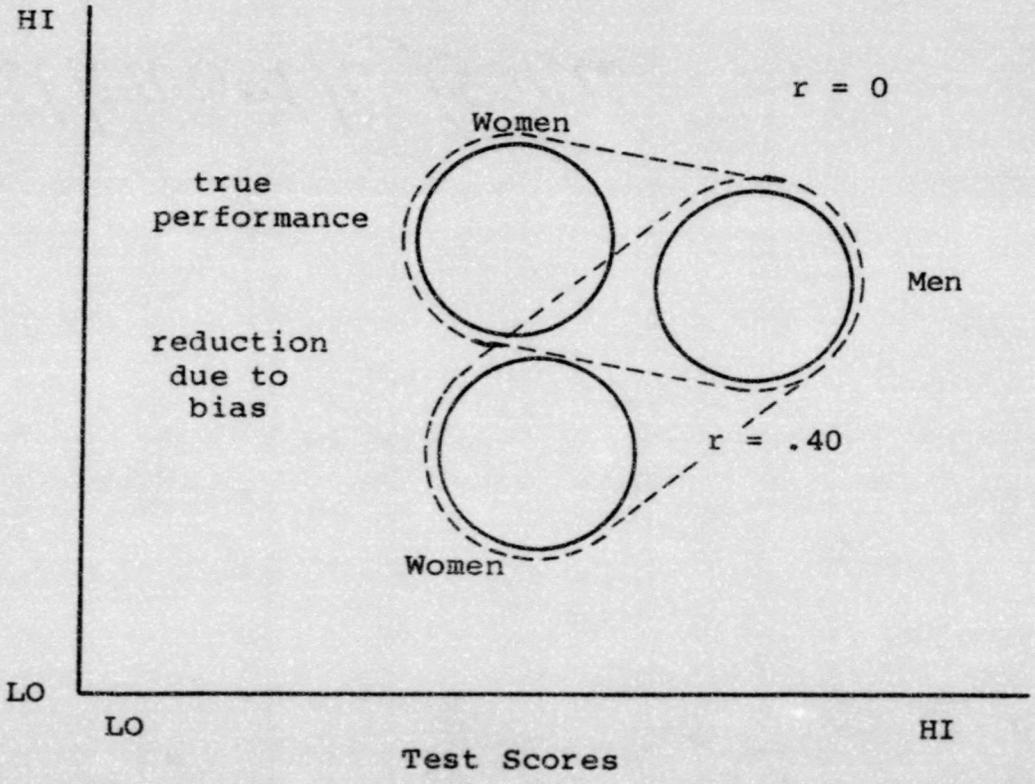
The gender bias in ratings found in the present study indicates that studies using physically demanding task

performance ratings as criterion measures must examine subgroup validities. If subgroup differences are not examined, use of the test may result in inadvertent discrimination against women. Subgroup differences should be examined to determine whether differential or single-group validity exists and to determine the appropriate predictor situation, i.e., whether selection decisions should be based on one regression equation or separate equations for each subgroup. The Uniform Guidelines on Selection Procedures (1978) and the APA Standards (1974) recommend that test users investigate differences in criterion-related validity for relevant subgroups, i.e., gender and race. Users are warned to conduct investigations of differential and single-group validity only when it is technically feasible, i.e., when subgroup sample sizes are large enough for reliable comparisons and relevant unbiased criteria are available. Although research indicates that differential and single-group validity rarely exist in well-controlled studies (Cascio, 1982), most of this research has examined racial differences in cognitive abilities (Arvey, 1979). The research examining differences between men and women in the area of physical abilities is not as conclusive (Arvey).

Before conducting an investigation in differential or single-group validity, possible differences in predictor and criterion scores should be examined. In the area of

physical abilities testing, particularly strength testing, women typically score lower than men, and small individuals tend to score lower than large individuals (Fleishman, 1964). These test differences are real and do not constitute bias. The problem arises when criterion ratings of women and small individuals are below their true performance. Because the physical abilities scores of women and small individuals are usually lower than those of men and large individuals, biased criterion ratings could result in a spurious validity coefficient which is driven by gender and body size bias (see Figure 1).

Figure 1. Comparison of true and gender biased physically demanding task performance ratings.



The results of the present study indicate that ratings of physically demanding task performance may indeed contain gender bias. Before using physically demanding task performance ratings in validity studies, users should examine the ratings for gender bias by statistically comparing mean ratings for men and women. If bias is found, the effects of gender should be statistically removed by partialling gender from ratings, raters should be trained to more accurately rate performance, and/or another criterion measure should be used.

Investigating potential bias in predictor and criterion scores is more practical than investigating differential or single-group validity. Investigations of differential and single-group validity require larger subgroup sample sizes than are typically available (Cascio, 1982). Furthermore, demonstrating a lack of differential or single-group validity does not assure fair use of the test. Mean differences in predictor and criterion scores must be considered to determine test fairness. Perhaps the most common approach used to investigate test fairness is to compare regression slopes and intercepts for relevant subgroups.

Given lower performance ratings for women found in the present study and the fact that women and small individuals score lower on physical abilities tests, four predictor situations are possible (Bartlett and O'Leary, 1969). In

the first situation, the test appears equally valid for both groups, women score lower on the selection test, and women receive proportionately lower job performance ratings than men (see Figure 2). In this case, a single regression line is appropriate and fair. However, the use of this selection device results in adverse impact against women. Though not illegal, according to the Uniform Guidelines on Selection Procedures (1978), employers should consider available alternatives with less adverse impact.

Figure 2. Valid test with adverse impact.

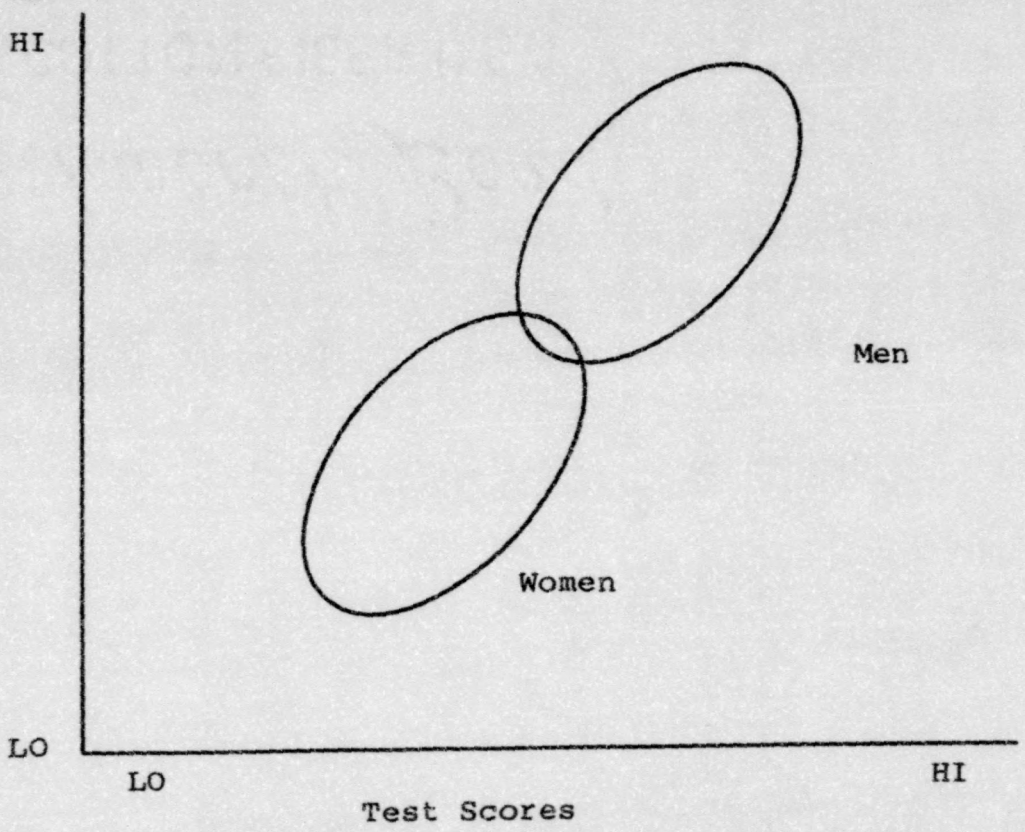
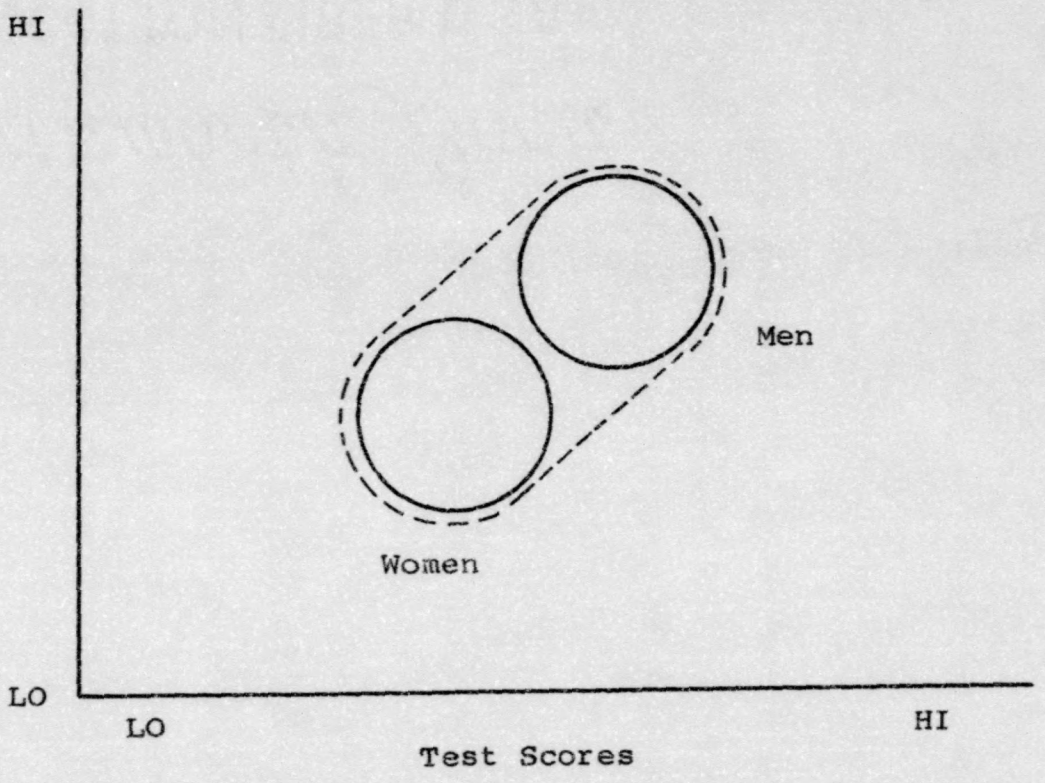


Figure 3 illustrates a situation in which the test is

valid for the group as a whole but is invalid for either subgroup. In this case, the selection test serves only as a crude predictor of gender, and using the test is the same as selecting applicants on the basis of gender. The use of this test is clearly illegal and unethical. However without examining validity for both subgroups, the test appears to be a valid predictor of job performance, and its use would result in inadvertent discrimination against women.

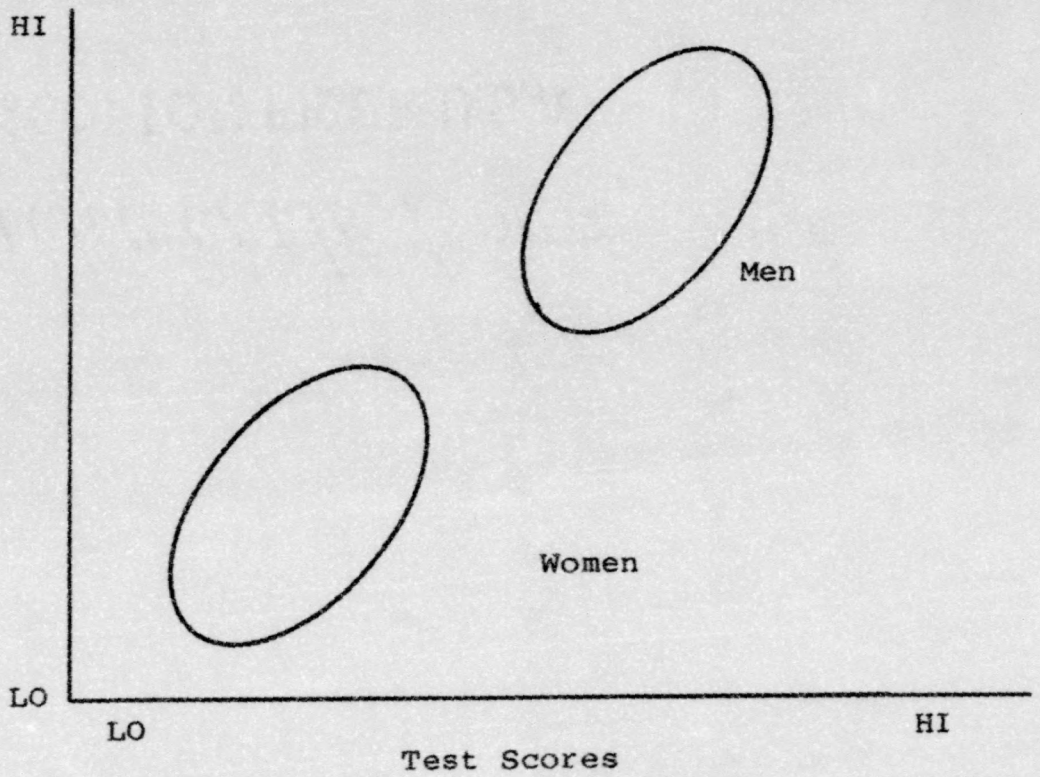
Figure 3. Valid test for group as a whole, invalid for each subgroup.



In the third case (see Figure 4), the test is valid for both subgroups, but women have lower predictor and criterion

scores. Here, using a single regression line results in an underprediction of job performance for men, an overprediction for women, and unfair discrimination against men. In this case, fairness is achieved by making predictions based on separate regression equations for each subgroup.

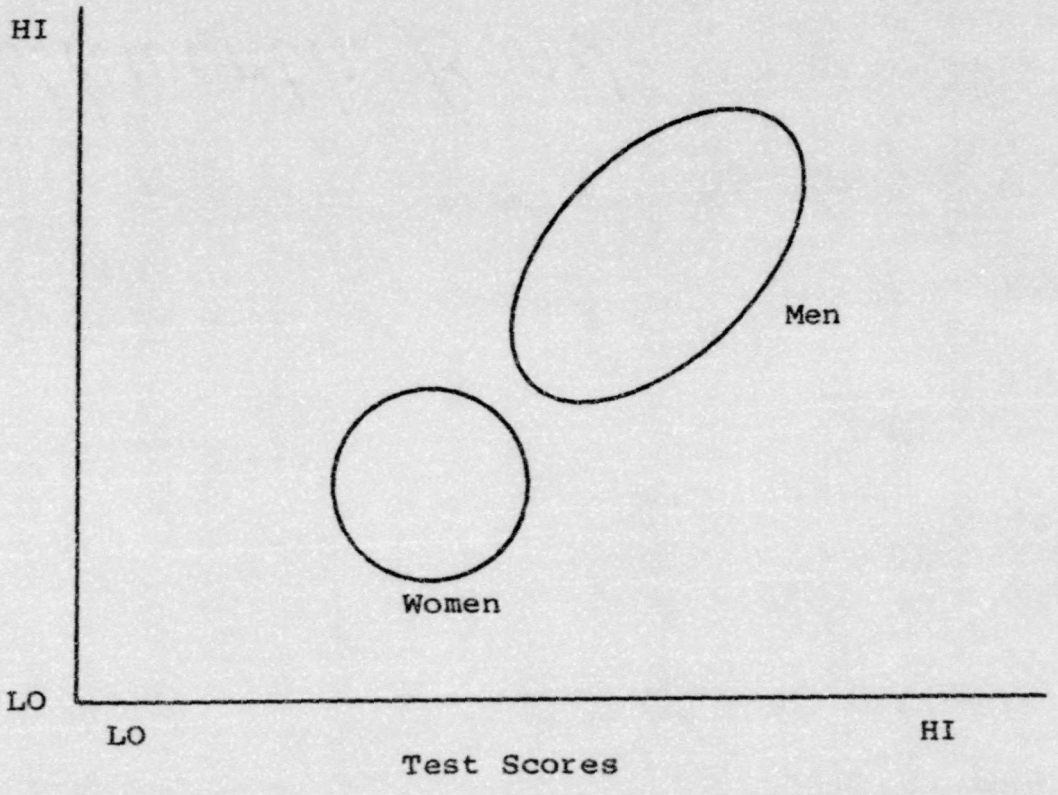
Figure 4. Equal validity for men and women, unequal predictor and criterion means.



In the final case, the predictor has no validity for women, and predictor and criterion means are different for the two subgroups (see Figure 5). The situation could be reversed so that the predictor is valid for women instead of

men. In this situation, the slopes of the two regression lines are not parallel and the intercepts intersect. Using a single regression equation will result in lower validity overall and lower predicted criterion scores. Because the selection test is not a valid predictor of job performance for women, it should not be used to select women. A regression equation for men can be calculated for this test, but another selection instrument must be used for women.

Figure 5. Validity for men only, unequal predictor and criterion means.



A selection test validated against gender biased performance ratings is likely to result in one of the four predictor situations described above. In the worst case

(Figure 3), the test would unknowingly be used to discriminate against applicants on the basis of gender. In the fairest case (Figure 2), the selection test would be both useful and fair but result in adverse impact. In the other two cases (Figures 4 and 5), the use of a single regression equation will lower the apparent validity of the test. Given gender biased performance ratings found in the present study, distribution of predictor and criterion scores for men and women must be examined to determine the appropriate predictor situation. Selection decisions should then be based on the appropriate regression equation or equations to ensure fair use of the selection test.

Although the main effect for body size approached significance ($F(7,192) = 2.98$, $p < .086$), the hypothesis of a body size difference in physically demanding task performance ratings was not supported. The lack of support for this hypothesis may be due to (a) the fact that body size does not influence physically demanding task performance ratings or (b) the weak manipulation of body size in the present study.

If body size does not influence physically demanding task performance ratings, spurious within groups validity coefficients based on body size can not be obtained for physical abilities tests. If significant within groups validity coefficients are obtained, it can be assumed that any variability in ratings within male and female subgroups

is random or due to some contaminating variable other than body size (e.g. experience).

The lack of a significant main effect for body size may be due to the weak manipulation of body size in the present study. Although the experimenter, who is somewhat above average in height (66 inches) and below average in weight (124 pounds), appeared with each actor during the instruction section of each videotape, on tape the actors did not visually appear strikingly larger or smaller than the experimenter.

Participants knew the actor's size when viewing the videotape and when rating performance, as demonstrated by their responses to the actor's height and weight questions (see Table 5), however this knowledge may have been totally derived from the dialogue and not from the visual cues pertaining to size. In other words, participants responded intellectually to the actor's size, but the perception of the actor as above or below average in size was not a salient factor when observing and rating performance. The same videotapes of the large and small men without the verbal height and weight information were used in a follow-up study to determine the impact of the visual cues of body size. Although a significant difference was found when participants were asked to estimate the height ($\underline{F}(1,76) = 9.02$, $\underline{p} < .003$) and weight ($\underline{F}(1,74) = 6.50$, $\underline{p} < .013$) of the two men, these differences were quite small and of no

practical significance (see Table 8). Thus, the size of the actor was not salient when observing his or her performance. Because of the weak size manipulation in the present study, the issue of the effects of body size on ratings of physically demanding task performance remains unresolved.

Table 8
Means and standard deviations of estimated height, in inches, and weight, in pounds, of large and small men in follow-up study.

	Estimated Height		Estimated Weight	
Actor	X	SD	X	SD
Large Man	70.9	1.89	172.6	15.93
Small Man	69.6	2.05	164.3	12.35

The data did not support the hypotheses of a gender or body size by time of rating interaction. In other words, there were no bias differences in ratings obtained immediately after viewing objective performance and those obtained one week later. It may be that one week is not enough time delay to effectively evaluate the effects of time on bias in ratings. Most supervisors rate their employees on an annual basis. They are often quite removed from the daily activities of their employees and are unable to observe all relevant job performance behaviors during that time (Borman, 1978). Thus, observing all relevant behavior for one person one week prior to rating his or her performance may not be a close simulation of the rating

situation in the real world. Further research must be conducted to determine the relationship between biased ratings and the time delay between observation and evaluation.

A gender bias in physically demanding task performance ratings has important personnel implications other than test validation. Performance appraisals often serve as bases for personnel decisions including promotions, training, transfers, termination, and salary increases. A gender bias in ratings that are widely used in personnel decisions could operate to deny women access to physically demanding jobs, training for those jobs, promotions and salary increases within those jobs, etc. One way to reduce gender bias in ratings and to increase rating accuracy is to train the rater.

Rater training programs have traditionally focused on the elimination of systematic error (e.g. halo, central tendency, leniency-severity, etc.) (Cascio, 1982) and are called rater error training (RET). However, these programs typically have only short-term effects (Ivancevich, 1979; Bernardin, 1978). More successful rater training programs have focused on training raters to more accurately rate behavior. Rater accuracy training (RAT) programs have consisted of one of two types. Performance dimension training (PDimT) informs raters of the performance dimensions to be rated. Performance standard training

(PStandT) teaches raters to judge performance against a desirable standard. RAT programs that incorporate both PDimT and PStandT are the most successful at increasing rater accuracy (Smith, 1986). Most rater training research has focused on eliminating rating error in nonphysically demanding positions (e.g. supervisors, teachers, interviewees) (Smith). Future research should be focused on eliminating gender bias in ratings of physically demanding task performance.

In summary, the gender differences in ratings obtained in the present study indicate that performance ratings of physically demanding jobs may be contaminated by sex role stereotype bias. Ratings are widely used as job performance measures in validity studies and as a basis for many personnel decisions (e.g. retention, promotion, pay increases, etc.). Because of potential gender bias in these ratings, women may be placed at the bottom of the rating distribution which may result in unfair use of selection tests and unfair personnel decisions. Not only do unfair use of selection tests and unfair personnel decisions work against individuals who desire physically demanding jobs by denying them access to these jobs, but they also hinder the effectiveness of organizations through decreased productivity and lower job satisfaction.

References

Abramson, P. R., Goldberg, P. A., Greenberg, 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.

American Psychological Association. (1974). Standards for educational and psychological tests. Washington, DC: Author.

Arvey, R. D. (1979). Fairness in selecting employees. Reading, MA: Addison-Wesley Publishing Company.

Arvey, R. D. & Begalla, M. E. (1975). Analyzing the homemaker job using the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 60, 513-517.

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.

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

Cascio, W. F. (1982). Applied Psychology in personnel management (2nd ed.). Reston, VA: Reston Publishing Company, Inc.

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

Darley, J. M. & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. Journal of Personality and Social Psychology, 44, 20-33.

Deaux, K. & Emswiller, T. (1974). Explanations of successful performance on sex-linked tasks: What is skill for the male is luck for the female. Journal of Personality and Social Psychology, 29, 80-85.

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.

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

Etaugh, C. & Brown, B. (1975). Perceiving the causes of success and failure of male and female performers. Developmental Psychology, 11, 103.

Feldman, J. M. (1981). Potential behavioral consequences of attributions of locus of control. Journal of Applied Psychology, 66, 63-68.

Feldman-Summers, S. & Kiesler, S. B. (1974). Those who are number two try harder: The effect of sex on attribution of causality. Journal of Personality and Social Psychology, 30, 846-855.

Fleishman, E. A. (1964). The structure and measurement of physical fitness. Englewood Cliffs, NJ: Prentice-Hall,

Inc.

Frank, F. D. & Drucker, J. (1977). The influence of evaluatee's sex on evaluation of a response of a managerial selection instrument. Sex Roles, 3, 59-64.

Gutek, 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.

Hogan, J. A model of physical performance for occupational tasks.

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

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

Maher, P. T. (1984). Police physical ability tests: Can they ever be valid? Public Personnel Management Journal, 13, 173-183.

Mobley, W. H. (1982). Supervisor and employee race and sex effects on performance appraisals: A field study of adverse impact and generalizability. Academy of Management Journal, 25, 598-606.

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

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.

Ross, M. & Fletcher, G. J. O. (1985). Attribution and social perception. In G. Lindsey & E. Aronson (Eds.), The handbook of social psychology. Vol. 2 (pp. 73-122). New York: Random House.

Schneier, C. E. & Beusse, W. E. (1980). The impact of sex and time in grade on management rating in the public sector: Prospects for the Civil Service Reform Act. Proceedings of the Academy of Management, 40, 329-333.

Smith, D. E. (1986). Training programs for performance

appraisal: A review. Academy of Management Review, 11, 22-40.

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.

Uniform Guidelines on Selection Procedures (1978). Federal register, 43, 38290-38315.

U.S. Department of Health, Education, and Welfare. (1979). Weight and height of adults 18-74 years of age: United States, 1971-79 (DHEW Publication No. PHS 79-1659). Hyattsville, MD: National Center for Health Statistics.

Appendix A

Instructions for Experimental and Control Delayed Rating Conditions

I'm conducting a study in which today I'll be asking you to view a videotape. Next week, I'll return to ask you to complete a questionnaire. If you don't want to participate, you may leave, wait in the hall, or sit quietly at your desk for the next 15-20 minutes. Please do not disturb those who choose to participate.

I'm conducting a study of the relationship between one's prior work history and performance ratings. Most performance appraisals are completed by supervisors who rate their employees' performances. This is the type of appraisal I'm interested in.

What I'd like for you to do is to assume that you are the 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 you will observe shows 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 compare how the applicant appears at the beginning of the session and at the conclusion of 30 minutes of continuous work. You will now carefully view the tape, after which you will be asked to rate the applicant's performance on three characteristics: (a) the amount of effort exerted, (b) the degree of fatigue that is apparent, and (c) overall performance. Today you will view the tape only. One week from today I will return and ask you to rate the applicant's performance. Therefore, you need to pay close attention to the tape.

Before we begin and to help you get a feel for how physically demanding the 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 knees like this .... (Demonstrate proper lifting technique. Wait for each participant, in an organized fashion, to lift the bag of feed.) Are there any questions? (After answering any questions, show the videotape. When the break in the tape occurs, say "You'll notice that the film has been cut here. We're now observing the last three minutes of the test.")

I'll be back next week for you to rate the applicant's performance. You will not have another opportunity to view the videotape so try to remember as much about the applicant and the applicant's performance as possible. Keep in mind that you are rating the applicant's performance on the amount of effort exerted, the degree of fatigue that is apparent, and overall performance. Also, I'd prefer that you didn't discuss the study with anyone until you have completed the ratings next week.

Last week you viewed a videotape of an applicant moving bags. You will now complete a questionnaire and rate that performance.

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

three forms you will complete today. What I'd like for you to do is to describe the most physically 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 completed the Work History Questionnaire, collect the questionnaires while handing out the Rating Form.) I'm passing out the Rating Form now. Do not begin to complete this form until I've gone over the instructions. Write your name in the upper right corner. Now I'd like for you to rate the performance of the applicant in the videotape. Read over the entire form before making any ratings. Then go back and rate each item carefully. Pay close attention to the verbal descriptions on the scale. Indicate your responses by placing an 'X' on the line closest to the answer which best reflects your opinion. Please be as accurate as possible when making these ratings. As you finish rating, please remain seated and don't communicate with others. Are there any questions?

(When everyone has completed the Rating Form, collect them while handing out the Supplemental Rating Form.) Write

your name in the upper right corner. Indicate your response by circling the one letter, filling in the blanks, or placing an 'X' on the line closest to the answer which best reflects your opinion. Please remain seated when you have completed this form, and don't communicate with others. Are there any questions?

(When the participants have completed the Supplemental Rating Form, collect all forms, and debrief the participants.)

Appendix B

Instructions for Experimental and Control Immediate Rating Conditions

I'm conducting a study in which today I'll be asking you to view a videotape and complete a questionnaire. If you do not want to participate, you may leave, wait in the hall, or sit quietly at your desk for the next 30-35 minutes. Please do not disturb those who chose to participate.

I'm conducting a study of the relationship between one's prior work history and performance ratings. Most performance appraisals are completed by supervisors who rate their employees' performances. This is the type of appraisal I'm interested in.

What I'd like for you to do is to assume that you are the 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 mot 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 you will observe shows 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 compare how the applicant appears at the beginning of the session and at the conclusion of 30 minutes of continuous work. You will now carefully view the tape, after which 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 knees like this .... (Demonstrate proper lifting technique. Wait for

each participant, in an organized fashion, to lift the bag of feed.) Are there any questions? (After answering any questions, show the videotape. When the break in the tape occurs, say "You'll notice that the film has been cut here. We're now observing the last three minutes of the test.")

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 three forms you will complete today. What I'd like for you to do is to describe the most physically 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 completed the Work History Questionnaire, collect the questionnaires while handing out the Rating Form.) I'm passing out the Rating Form now. Do not begin to complete this form until I've gone over the instructions. Write your name in the upper right corner. Now I'd like for you to rate the performance of the

applicant in the videotape. Read over the entire form before making any ratings. Then go back and rate each item carefully. Pay close attention to the verbal descriptions on each scale. Indicate your responses by placing an 'X' on the line closest to the answer which best reflects your opinion. Please be as accurate as possible when making these ratings. As you finish rating, please remain seated, and don't communicate with others. Are there any questions?

(When everyone has completed the Rating Form, collect them while handing out the Supplemental Rating Form.) Write your name in the upper right corner. Indicate your response by circling the one letter, filling in the blanks, or placing an 'X' on the line closest to the answer which best reflects your opinion. Please remain seated when you have completed this form, and don't communicate with others. Are there any questions?

(When the participants have completed the Supplemental Rating Form, collect all forms, and debrief the participants.)

Appendix C

Work History Questionnaire

NAME.....

Work History Questionnaire

Describe the most physically demanding work you have ever done. Include volunteer work, housework, farm labor, & 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 and year)
   From: To:

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

4. On 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?

6. If so, how heavy, in pounds, were the objects you lifted?

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

8. Did you lift the objects over your head?

9. How many feet did you move the objects?

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

Appendix D
Rating Form

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 item. Answer by placing an 'X' on the line closest to the answer which best reflects your opinion. Be as accurate as possible.

EFFORT

1. While performing the task, the applicant appeared to be under

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

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

very low effort	low effort	average	some effort	very high effort

3. To complete the task, the applicant seemed to struggle

a great deal	somewhat	average	a little	not at all

FATIGUE

4. After performing the task, the applicant appeared to be

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

5. 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 6 hours

6. 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 an 8 hour shift would this applicant need to take this kind of break?

frequently	occasionally	sometimes	rarely	never

PERFORMANCE

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

300	240	180	120	60

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

strongly agree	agree	neutral	disagree	strongly disagree

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

strongly not recommend	not recommend	neutral	recommend	strongly recommend

10. Overall, the applicant's performance was

superior	above average	average	below average	poor

Appendix E
Supplemental Rating Form

NAME.....

Supplemental Rating Form

Answer as accurately as possible by circling the one letter, filling in the blanks, or placing an 'X' on the line closest to the answer which best reflects your opinion.

1. The applicant performed the way he or she did because of
   a. luck.
   b. effort.
   c. ability.
   d. the difficulty level of the task.

2. Based on my recollection, the applicant is ..... feet ..... inches tall and weighs ..... pounds.

3. Compared to the average height and weight of adult males and females, the applicant is ..... for his or her sex.

---

above average average below average