RACIAL BIAS AS A CONTAMINANT OF PERFORMANCE EVALUATIONS

JAMES L. OUTTZ

Research Report No. 15

February, 1977

This research was partially supported by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research under Contract No. N00014-75-C-0884, Contract Authority Identification Number, NR 151-375, Benjamin Schneider and C. J. Bartlett, Principal Investigators.

Reproduction in whole or part is permitted for any purpose of the United States Government. Approved for public release; distribution unlimited.
Performance evaluation is an integral part of the organizational process. Most organizations have a more or less formal system of procedures for evaluating employee performance. Often, the validity of such procedures is not determined. This can present significant problems when various components of the organizational process, such as selection or training, depend upon accurate appraisal of job performance.
Supervisory ratings are the most common method of performance evaluation. In using such ratings, one employee (usually a supervisor) is required to rate another employee on several traits or performance dimensions. The pitfalls inherent in the use of such criteria are well known. However, relatively little research has been done on racial bias as a source of contamination in performance ratings. The primary purpose of this research was to assess the degree to which rating behavior is affected by the race of rater and ratee.

Three primary hypotheses were tested: (1) Evaluations of job performance are related to the race of rater and ratee; (2) In the absence of relevant information perceived attitude similarity is related to the race of the individuals involved; and (3) The effects of racial bias on performance ratings will depend upon the cognitive balance between rater, ratee, and the ratee’s performance.

Eight hypothetical college instructors were evaluated by 174 white and 129 black college students. Each instructor's teaching performance was described using statements of predetermined favorability. Three levels of teaching performance, good, average, and poor were described. The race of the instructors was manipulated through the use of photographs. Each subject evaluated every instructor in terms of (a) quality of performance, (b) promotability, (c) agreement with his teaching philosophy, (d) agreement with his philosophy about students and (e) expected change in future performance.

The hypotheses were tested using correlational analyses. Race of rater and ratee was not related to the performance evaluations of good instructors. Where teaching performance was described as poor, white raters favored white instructors over black instructors in terms of quality of performance. Where teaching performance was described as average, white raters favored average black instructors over white instructors in terms of promotability. Black raters favored average black instructors over average white instructors in terms of quality of performance and promotability. It was deduced that performance ratings were based primarily on teaching performance, but, when teaching performance was not exceptional, rater-ratee racial similarity was a factor as well.

When teaching performance was described as very good, there was no significant relationship between rater-ratee racial similarity and agreement with the ratee's beliefs. However, white raters indicated that they would agree more with the beliefs of a poor white instructor than a poor black instructor. Black raters indicated that they would agree more with the beliefs of an average black instructor than an average white instructor. It was concluded that perceived belief similarity was based more on teaching performance than race, but, when teaching performance was not exceptional, rater-ratee racial similarity was a factor. Manipulation of the independent variable was not sufficient to determine the effects of cognitive balance on rating behavior.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Criterion Contamination</td>
<td>1</td>
</tr>
<tr>
<td>Leniency in Ratings</td>
<td>2</td>
</tr>
<tr>
<td>Halo</td>
<td>4</td>
</tr>
<tr>
<td>Ethnic Group Membership as a Contaminant of Performance Evaluations</td>
<td>4</td>
</tr>
<tr>
<td>Ethnic Group Membership as it Affects Non-Supervisory Performance Ratings</td>
<td>10</td>
</tr>
<tr>
<td>Criterion Bias in Test Validation</td>
<td>11</td>
</tr>
<tr>
<td>An Experimental Study</td>
<td>12</td>
</tr>
<tr>
<td>Processes Underlying Racial Bias in Criterion Ratings</td>
<td>16</td>
</tr>
<tr>
<td>Balance Theory</td>
<td>17</td>
</tr>
<tr>
<td>Purpose of the Present Investigation</td>
<td>18</td>
</tr>
<tr>
<td>II. METHOD</td>
<td>20</td>
</tr>
<tr>
<td>Subjects</td>
<td>20</td>
</tr>
<tr>
<td>Procedure</td>
<td>20</td>
</tr>
<tr>
<td>Method of Data Collection</td>
<td>26</td>
</tr>
<tr>
<td>Debriefing</td>
<td>26</td>
</tr>
<tr>
<td>Research Hypotheses</td>
<td>27</td>
</tr>
<tr>
<td>III. RESULTS</td>
<td>30</td>
</tr>
<tr>
<td>Primary Findings and Post Hoc Analyses</td>
<td>30</td>
</tr>
<tr>
<td>Manipulation of the Independent Variables</td>
<td>30</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. College Experience of the Subjects</td>
<td>21</td>
</tr>
<tr>
<td>2. Order of Presentation of Information</td>
<td>24</td>
</tr>
<tr>
<td>3. Mean Performance Ratings of Good, Average and Poor Instructors</td>
<td>31</td>
</tr>
<tr>
<td>4. Summary Table of the Performance Ratings of White Instructors by White Males</td>
<td>34</td>
</tr>
<tr>
<td>5. Summary Table of the Performance Ratings of Black Instructors by Black Males</td>
<td>36</td>
</tr>
<tr>
<td>6. Zero Order Correlations Between Level of Performance and Performance Ratings</td>
<td>37</td>
</tr>
<tr>
<td>7. Second Order Partial Correlations Between Race of Ratee and Performance Ratings by Race of Rater</td>
<td>39</td>
</tr>
<tr>
<td>8. Second Order Partial Correlations Between Race of the Instructor and Ratings of Belief Similarity by Race of Rater</td>
<td>41</td>
</tr>
<tr>
<td>9. Average Rating of Expected Change by Race and Degree of Cognitive Balance</td>
<td>42</td>
</tr>
<tr>
<td>10. Partial Correlations Between Instructor's Race and Performance Ratings by Level of Performance</td>
<td>44</td>
</tr>
<tr>
<td>11. Mean Ratings of Performance by Race of Rater and Ratee</td>
<td>45</td>
</tr>
<tr>
<td>12. Partial Correlations Between Race of Ratee and Ratings of Belief Similarity for Each Level of Teaching Performance</td>
<td>46</td>
</tr>
<tr>
<td>13. Mean Ratings of Belief Similarity by Race of Rater and Ratee</td>
<td>47</td>
</tr>
<tr>
<td>14. What Students Thought Most Important for an Instructor to Accomplish</td>
<td>49</td>
</tr>
<tr>
<td>15. Students' Thoughts as They Evaluated the Instructors</td>
<td>50</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Analysis of Variance Repeated Measurements Design</td>
<td>33</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

Job performance criteria are rules or standards by which work performance can be judged. Examples of such criteria include, quality of work, quantity of work, absenteeism, promotability, and supervisors' ratings. When appropriately developed, and applied, criterion measures are useful indices of job performance. Unfortunately most job performance criteria are poorly developed and inequitably applied. As a consequence, selection and training programs often contain the same faults as the criteria on which they were developed.

Criterion Contamination

Although most researchers recognize that contaminated criterion measures can significantly affect predictive validity, the problem has been virtually ignored. This is especially true where employment discrimination is concerned. Most of the research in this area has centered around the question of differential predictability of minorities and non-minorities, (Bartlett & O'Leary, 1969; Bray & Moses, 1972; Cleary, 1968; Farr, O'Leary & Bartlett, 1971; Kirkpatrick, et al., 1968; Lefkowitz, 1972; Lopez, 1966). In this research little attention was given to criterion contamination.

Criterion contamination can be defined as the degree to which the measures actually used to determine job success (e.g. supervisory ratings) contain elements that are unrelated to true measures of job success.
(Blum & Naylor, 1968). One of the components of criterion contamination is systematic error or criterion bias. Criterion bias refers to those components of an actual criterion which are, (a) unrelated to job success and, (b) correlated with one or more variables such as rater leniency or opportunity bias.

Prior to focusing directly on the issue of racial bias as a contaminant of criterion measures, it may be useful to consider ways in which criterion contamination has been conceptualized in the past. Kipnis (1960), in discussing the lack of attention given to bias in criterion ratings, outlined some of the factors which may contribute to it. He proposed that a supervisor may systematically err in rating a subordinate by rating too leniently, too harshly, or by allowing a few performance dimensions to unduly affect evaluation of other performance dimensions.

**Leniency in Ratings**

Some of the factors which seem to influence rater leniency are propinquity between rater and ratee; the amount of pressure under which the supervisor works; the degree to which the supervisor is allowed to criticize a subordinate during the course of their interactions; and the type of rating scale employed.

**Propinquity.** Physical distance between individuals, such as employees, dictates whether they develop friendships (Byrne & Buehler, 1955; Caplow & Forman, 1950; Festinger, Schacter & Bach, 1950; Kipnis, 1957). This suggests that the closer, in physical terms, an employee is to his or her supervisor, the greater the likelihood of their becoming friends. This may in turn affect the employee's performance rating. In the industrial setting, factors such as the physical arrangement of work space or the assignment of work tasks, may cause leniency in performance ratings.
**Social setting.** Another factor which affects rater leniency is the social setting within which performance evaluations are made. Changes in interpersonal structure within the work setting can produce changes in the degree to which group members like each other. Such changes could influence the performance evaluations of the employees involved.

**Pressure.** The degree of pressure a supervisor receives from the social setting may affect his or her ratings of subordinates. Research by Murray (1933) and Feshbach and Singer (1957) indicates that stress brought about by pressure will decrease the level of positive sentiment between individuals. The greater the stress, the more likely the rater is to project maliciousness and aggression onto the person being rated. Rigidly enforced personnel policies or outside pressures such as EEOC guidelines could produce stressful pressure resulting in systematic bias in criterion ratings.

**Criticism.** Research on rater leniency also indicates that individuals who are restrained from criticizing others tend to develop less friendly feelings toward them (Thibaut & Coules, 1952; Pepitone & Kleiner, 1957). Spector (1954) found that blocking rater criticism of a subordinate’s performance tended to result in less lenient evaluations. He found that the most lenient ratings were given by supervisors who were allowed to criticize the ratee.

**The rating scale.** The rating scale itself can influence the degree to which leniency occurs. It has been well documented that rating scales differ in the degree to which they minimize rater leniency (Campbell, Dunnette, Arvey & Hallervik, 1973; Madden & Bourdon, 1964; Sharon & Bartlett, 1969). Madden and Bourdon (1964) constructed nine variations of mean occupation evaluation ratings. The results showed that different rating scale formats produced significant differences
in evaluations. Campbell et al. (1973) developed behaviorally anchored rating scales for managerial performance. The behaviorally anchored scales were compared with traditional Likert type scales. The results showed that the Likert type scales were more susceptible to leniency error. Sharon and Bartlett (1969) assessed the degree to which a forced choice rating scale minimized leniency error in comparison with a graphic rating scale. Leniency was defined as a favorable shift in mean ratings from one rating condition to another. They found that the forced choice scale yielded significantly less leniency error.

Halo

There is evidence which indicates that a limited class of behaviors can influence ratings of other aspects of job performance. That is, a particular trait can influence the evaluation of other aspects of job performance. This may explain differences often found between supervisory ratings and less judgemental measures of ratee performance.

The research literature indicates that performance ratings contain some elements of bias. This bias has often been conceptualized as a tendency on the part of raters to allow attitudes toward a subordinate to influence evaluations of the subordinate's job performance. Little research has been done however to determine whether characteristics of the ratee such as ethnic group membership or the ratee's sex are related to evaluations of job performance.

Ethnic Group Membership as a Contaminant of Performance Ratings

In a study more directly relevant to the topic of this thesis, Bass and Turner (1973) studied the significance of ethnic group membership in relationships among criteria of job performance. Their purpose
was to assess the degree of racial discrimination or systematic bias in criterion measures of black and white bank tellers. This particular sample provided the opportunity to study both non-judgemental and rating-type measures of job performance. Two hundred and forty-four part-time and 190 full-time tellers were studied. Thirty-two of the part-time tellers and forty-three full-time tellers were black. All of the tellers performed basically the same work tasks. The performance evaluations were obtained from 163 supervisors. Performance was rated on the basis of five job performance factors and overall effectiveness. The five factors were (1) customer relations, (2) ability to sell new accounts and services, (3) quality of work, (4) alertness to irregularities, and (5) cooperation with others. Each of the five factors was evaluated using a five point scale. Overall effectiveness was rated on a ten point scale. Approximately 84% of the employees were rated by two supervisors, allowing inter-rater reliabilities to be obtained. The median reliability for the six scales (five factors plus overall effectiveness) was .76.

In addition to supervisory ratings, four non-judgemental measures of job performance were obtained. These were (a) number of shortages obtained over a one month period, (b) number of overages obtained for the same period, (c) attendance (computed only for full-time tellers) and, (d) net salary increase for each teller since first coming to work (i.e. salary increase corrected for tenure). Measures of three control variables, tenure, age and race were also obtained. The results showed that white employees, both full-time and part-time, had higher mean scores on practically all of the criterion variables. Mean differences were significant with regard to customer relations, quality of work, alertness to irregularities, cooperation, overall effectiveness, number
of shortages, and adjusted salary increase. White employees were also older and more experienced. However, this does not seem to account for the differences in criterion ratings.

Since there were definite disparities between the mean ratings of the two groups, a question arose as to whether the race of the rater or ratee influenced rating behavior. Intercorrelations were computed between all of the variables. Race was significantly correlated with both supervisory ratings and objective measures of job performance. Correlations were also computed between race and the criterion measures with tenure and age partialled out. It should be recalled that the mean age and job experience of the white employees were higher than that of blacks. This might have lead to the conclusion that the relationships between race and the criterion variables were simply artifacts of differential job experience. The partial correlations indicated that this was not the case. There were four significant correlations before age and tenure were statistically controlled. After these variables had been partialled out, two of the four relationships (percent of the time worked and number of shortages) remained significant. In the case of part-time employees, four of the original seven correlations (quality of work, overall effectiveness, number of shortages and adjusted salary increase) were significant after age and tenure were partialled out.

Control for the effects of age and tenure reduced but did not eliminate the relationship between race and measures of job performance.

The fact that blacks received significantly lower performance ratings than whites does not necessarily mean that they were discriminated against. It is possible that the black employees as a group were poorer performers than white employees. In any event, the major conclusion drawn by Bass and Turner was that the supervisors tended to use
less judgemental data in evaluating black tellers but did not do so for whites. Their conclusion was based upon the finding that supervisory ratings of black employees were more consistent with less judgemental indices of job performance (e.g. number of shortages and overages) than ratings of white employees. Another explanation for Bass and Turner's results is that the supervisors (all white) felt psychologically distant from black employees.

Regardless of whether there were differences in the factors used to evaluate the tellers or differences in psychological distance, the results of Bass and Turner's study indicate that it would be difficult to develop predictors of the bank tellers' performance. If supervisory ratings were used as the criterion, a given predictor would probably not show the same degree of correlation with the performance ratings of blacks as it would with ratings of whites. The supervisors apparently used different standards for evaluating the two groups. If as Bass and Turner report, non-judgemental data (i.e. shortages and overages) was used more in evaluating black employees, a measure of numerical ability might show the highest degree of correlation with their performance ratings. Such a predictor would probably show little correlation with performance ratings of white employees since use of objective data was not emphasized in evaluating them. Personality or motivational factors were apparently used more in evaluating white employees. Therefore some sort of personality inventory might be more useful in predicting their performance ratings. In order to utilize the same predictor(s) for both groups, job performance standards would have to be applied more uniformly.

Flaugher, Campbell and Pike (1969) present data from a study designed to assess whether, (a) criterion ratings are affected by the race
of the rater and ratee, and (b) whether rater-ratee racial group membership affects the correlation between subjective and objective criterion measures. Their study was part of a larger investigation which dealt directly with validation of a predictor against a criterion categorized on the basis of the race of rater and ratee. In the 1969 study the authors developed a nine scale performance rating form with which to assess the job performance of 168 black and 296 white medical technicians. All of the technicians were employees of U. S. Veterans Administration Hospitals throughout the country. Each technician was rated by at least two supervisors. The average rating was used for analysis. The nine rating scales were developed from interviews with incumbents. In addition to supervisory ratings, a job knowledge test was developed as a more objective measure of job performance. White technicians received slightly higher ratings on all of the scales except flexibility. White technicians also received higher scores on the job knowledge test. A Wilcoxon matched pairs signed-ranks test was used to test the difference between the two groups on all of the ratings taken collectively. The difference was significant at the .01 level. A t-test indicated that the mean difference between groups on the job knowledge test was significant at the .01 level. The researchers further sub-grouped their data by racial combination of rater and ratee. When analysed in this manner, the results showed that black technicians rated by black supervisors received higher ratings than blacks rated by whites. This was true for all of the scales and the job knowledge test. There are several hypotheses as to why this might have occurred.

One hypothesis is that black supervisors were more lenient toward technicians of their own race. This hypothesis seems tenable even though ratings of technicians rated by black supervisors were highly
correlated with scores on the job knowledge test. A second hypothesis is that black supervisors were generally more lenient than white supervisors. This is definitely a possibility. There is little certainty based upon the data that either group actually performed better than the other. One way to solve this problem would be to control the characteristics of the ratees in such a manner that each rater receives the same information about the person to be rated. Still a third hypothesis is that the two groups of supervisors placed differential weight on various criterion components in rating black and white subordinates. This hypothesis seems tenable considering the correlations between performance ratings and scores on the job knowledge test. The correlations were higher for black technicians rated by black supervisors than for white technicians rated by black supervisors. It appears that black supervisors rated black employees on the basis of criterion elements similar to those measured by the job knowledge test. They apparently did not do so in rating white employees. A fourth hypothesis is that white supervisors were more objective when rating black tellers. This hypothesis also seems tenable given that the correlations of performance ratings with test scores were higher for blacks rated by whites than for whites rated by whites. This pattern did not exist for black raters. Their ratings of white subordinates appeared to be less objective than their ratings of black subordinates.

All of the hypotheses above could explain the results of the Flaugher study. The major obstacle to eliminating these hypotheses is control of the quality of ratee performance. If this were accomplished, more valid estimates of rater bias would result.
Ethnic Group Membership As It Affects Non-Supervisory Performance Ratings

Obviously there are measures other than supervisory judgements which may be used to differentiate acceptable from unacceptable job performance. One such measure is the peer rating. Although they are not used as often as supervisory ratings, some research has been done regarding the effects of ethnic group membership on peer ratings. This research provides a glimpse of the extent to which race of the rater and ratee influence performance measures other than supervisory ratings.

Cox and Krombolz (1958) collected leadership peer ratings from 63 black and 469 white basic airmen at an air force base in Texas. Peer ratings were collected as the men began their third week of basic training. The men were assigned to various flights in the order in which they arrived for processing. Each flight consisted of about 60 airmen. The airmen rated each other on the basis of the following instructions:

Several airmen from your flight will be selected for positions such as squad leader, right guide, assistant flight leader and so forth. You are asked to rate each member of your flight as to how well you think he can do such a job. The selection of airmen for positions will be made on the basis of your ratings.

The ratings were made according to the following scale:

5 - Very well qualified for the position, one of the best in the flight
4 - A well qualified man for the position
3 - Would not do a bad job or a good one; about average
2 - Not well qualified for the job; would do a poor job at this time
1 - Poorly qualified for the job, would do a very poor job at this time

Ratings were sub grouped on the basis of the race of rater. Each man received two ratings. One was determined by averaging the ratings given by the blacks in his flight. The other rating was determined by
averaging the ratings given by the white members of his flight. An
analysis of variance was carried out on ratings given to blacks and to
whites. The results showed that there was an interaction between race
and flight. That is, blacks tended to give higher ratings to blacks,
but this was not true for all flights. By the same token, whites gave
higher ratings to whites but this was not true for every flight. Cox
and Krumboltz' data show that racial bias in criterion ratings may exist
even when peer ratings are used. However, the small number of airmen
participating in the study creates the possibility that friendships
between flight crew members not racial bias accounted for the findings.

Schmidt and Johnson (1973) studied the effects of race on peer
ratings in an industrial setting. Their results were supportive of the
data from the Cox and Krumboltz study, but their methodology was some-
what different. Instead of merely assessing the degree to which the
samples differed in their ratings of each other, Schmidt and Johnson
conducted sensitivity training specifically designed to eliminate racial
bias. Although there was still a tendency for raters to give higher
ratings to persons of their own race, the differences were not signifi-
cant.

There have been few studies reported in the literature assessing
the relationship between race and peer ratings. The evidence indicates
that higher ratings are given when rater and ratee belong to the same
race. More research must be done which eliminates rival hypotheses
before stronger conclusions can be drawn.

Criterion Bias in Test Validation

So far this discussion has dealt exclusively with racial group
membership as a moderator of the relationship between race and criterion
ratings. What effect does rater-ratee racial group membership have on
the correlation between a given predictor or predictors and criterion measures? Evidence bearing upon this question is scant. Campbell et al. (1970) conducted a concurrent validation study dealing directly with this issue. The purpose of the study was to validate a test against different criterion measures subgrouped according to various racial combinations of rater and ratee. The rater-ratee combinations were, (a) blacks rated by blacks; (b) blacks rated by whites; (c) whites rated by whites; and (d) whites rated by blacks. Correlations between supervisory ratings and each of nine aptitude tests were computed separately for each rater-ratee combination. The subjects were 198 black and 217 white medical technicians. The results showed that race of the rater and ratee may have affected the validity coefficients obtained. Only one of the predictors (arithmetic) showed a consistent correlation with supervisory ratings across all of the rater-ratee combinations.

An Experimental Study

The investigations reviewed thus far have dealt with both supervisory ratings and peer ratings. None of them however were conducted in a laboratory setting where ratee characteristics and rater observations could be controlled. This type of study would allow the researcher to assess racial bias which is experimentally induced. Hamer et al. (1974) conducted such a study. The purpose of their investigation was to assess whether performance ratings were free of rater bias when the quality of ratee performance was systematically controlled. Their subjects were thirty-six undergraduate business students. The subjects included nine black males, nine black females, nine white males, and nine white females. They were asked to play the role of a grocery store manager. In this capacity they were to screen for potential employment eight ratees who
had supposedly applied for the job of stock clerk. The eight hypothetical applicants consisted of two white males, two white females, two black males and two black females. The subjects were told that each hypothetical applicant had taken a three minute work sample examination as part of the screening process. The examination consisted of removing number ten cans from an opened case and placing them on grocery store shelves. The researchers had determined from personal experience that the average worker could stack 36 cans in a three minute period; a very good (high performance worker) could stack 48 cans; and a low performance worker could stack only 24 cans within the same three minute period. The eight job applicants or ratees were trained so that they all stacked the cans the same way. All of the ratees wore a light shirt, dark slacks and dark grocer's apron. A panel of graduate students agreed that the ratees looked similar in dress, method of task performance (i.e. use of both hands), quality of work and facial expression. The university media center designed a grocery store set and filmed the performance of the eight ratees. The film was shown to the 36 subjects in the following order:

1st - low performing black male
2nd - high performing white female
3rd - low performing black female
4th - high performing white male
5th - low performing white male
6th - high performing black female
7th - high performing black male
8th - low performing white female

Overall performance ratings of the eight ratees were obtained after the subjects had viewed the 24 minute film. The film was stopped after each three minute segment to allow the subjects time to write down comments about each applicant's performance. Applicants were rated on a 15 point scale with the following scale points:
(a) weak in overall performance, failed to produce acceptable amount of production
(b) performance periodically fell below that which I would expect from similar individuals in this task
(c) did a satisfactory job, was an average worker in his/her performance
(d) was a good worker, his/her performance was better than average
(e) an exceptionally good worker who will be equalled by few others

There were significant differences in the ratings based upon rater-ratee race and sex. A 2 x 2 x 2 x 2 x 2 design was used to test the effects of (a) rater's race (black/white), (b) rater's sex, (c) ratee's race (black/white), (d) ratee's sex and (e) ratee's performance level. An F-test for the main effect of level of performance was significant. High performance ratees were rated significantly higher than low performance ratees ($F = 70.45, \text{df} = 1/32, p < .001$).

There was also a significant main effect for sex of ratee. Female applicants were rated significantly higher than males ($F = 16.79, \text{df} = 1/32, p < .001$). This was especially true for good applicants. It was also found that black raters gave black applicants significantly higher ratings than white applicants (a mean of 8.50 to 7.72). White raters gave higher ratings to whites than to blacks (a mean of 8.27 to 7.80). An F-test for the main effect of rater-ratee race and sex was significant ($F = 4.71, \text{df} = 1/32, p < .05$). An interaction was found between race and level of performance. On overall task performance, low performing blacks and high performing blacks were rated similarly, 7.31 and 8.98 respectively. This difference was not significant. Low performing whites and high performing whites received quite dissimilar ratings, 10.36 and 6.63 respectively. Therefore, high performing whites were favored over high performing blacks. Omega square values indicated that although 30% of the variance in overall ratings was accounted for by performance level, an additional 23% was accounted for by different race-sex combinations.
The data from Hamner et al. appears to support the hypothesis that race-sex bias in performance ratings does occur. It would appear that raters tend to give higher ratings to individuals of their own race or sex. There are however some serious problems with the procedure employed by Hamner et al. which leave this hypothesis open to question. First, the instructional set resembled a selection procedure more than a performance evaluation procedure. Subjects rated the performance of fictitious job applicants not incumbents. The subjects were in essence screening applicants for hire using the work film as a predictor, as opposed to judging actual job performance. The study may have demonstrated bias in hiring preferences, but not bias in criterion measures. Secondly, differences in the ratings may have been due to contrast effects brought about by the order of presentation of the applicants. Wexley et al. (1972) found that students' ratings of job applicants (presented on video tape) were influenced by the ratings of preceding applicants. They found that when an applicant with high qualifications was preceded by applicants of high quality, he was rated lower than a similar applicant who was preceded by individuals of low quality. In the Hamner et al. study, high and low performing black males may have been rated similarly because high performing black males were preceded by high performing black females.

Aside from various methodological considerations, the Hamner study suffers from another serious short-coming. The authors made no attempt to measure the processes underlying variance in the dependent variable. No attempt was made to measure the psychological processes (i.e. prejudice, ethnocentrism, authoritarianism) which may have accounted for the differential evaluation of applicants on the basis of race. This criticism also applies to other research in this area.
Processes Underlying Racial Bias in Criterion Ratings

The research most relevant to the psychological processes underlying racial bias in criterion ratings comes from social psychological literature. Racial bias in performance evaluation seems to be similar to racial prejudice as defined by social psychologists. Rokeach, Smith and Evans (1960), propose that racial prejudice is in large part the result of perceived dissimilarity of beliefs. They hypothesize that a person does not reject another on the basis of that person's race alone, but rather rejects others who are perceived as having beliefs and values different from his own. Prejudice against blacks, for example, occurs because whites think blacks differ from them in important beliefs and values.

Research has been done to test Rokeach's theory (Byrne & Wong, 1962; Hendrick, Bixentine & Hawkins, 1971; Insko & Robinson, 1967; Johnson & Johnson, 1972; Rokeach & Mezei, 1966; Senn, 1971; Stein, 1966; Stein, Hardick, & Smith, 1965; Triandis, 1961; Triandis & Davis, 1965). The usual research procedure has been to measure subjects' interpersonal preference for others who supposedly differed from them racially, had different beliefs or both. Interpersonal preference was measured through the use of social distance scales ranging from "willingness to work with the stimulus person in an experiment" to "willingness to share an apartment with that person". The research has generally been supportive of the theory. When information about the stimulus persons' beliefs was made available, this information, not race was the primary determinant of attitudes toward the stimulus person. Research has also shown that race is the primary factor in determining attitudes toward a person where information about the person's beliefs is not available. As
an example, if a stimulus person is black, white subjects assume that the person's beliefs are different from theirs. If indeed there is a tendency for supervisors to rate subordinates of the same race more favorably, the reason might be perceived similarities in beliefs. The question is why should knowledge of an individual's beliefs affect evaluations of his or her work performance.

**Balance Theory**

There are several theories within the area of social psychology which appeal to central factors such as attitudes, expectations, knowledges and beliefs as basic elements in the explanation of behavior (Shaw & Costanza, 1970). Among these are the cognitive consistency theories. The basic premise of these theories is that inconsistent cognitions lead to behaviors directed toward the achievement of cognitive consistency. One of the most prominent consistency theories is Heider's balance theory (Heider, 1968).

Heider proposes that within the context of interpersonal relations, entities are seen as forming units. The relationships which exist between entities make a unit balanced or unbalanced. Balance theory is concerned with the sentiments (affective reactions) of a person "p" toward a person "o", and an impersonal object "x". According to the theory, separate entities such as "p", "o" and "x" constitute a unit when they are seen as belonging together. As an example, if person "p" likes both person "o" and an entity "x", and "o" produced "x", then the triad is said to be balanced. On the other hand, if "p" dislikes "o" but likes "x", and "o" produced "x", then the triad is unbalanced. Heider proposes that individuals tend to prefer balanced over unbalanced
situations because unbalanced situations create tension. Given a particular triad p-o-x, the tendency toward balance induces relations that result in a balanced state. If for example there exists a situation where "p" dislikes "o" but likes "x", and "o" produced "x", then there is a tendency for "p" to either change his sentiment toward "o" or change the positive relation of "o" produced "x". If a situation exists where a supervisor "p" dislikes a subordinate "o" but values good job performance "x", while at the same time perceiving that "o" produced "x", the triad would be unbalanced. One way to reduce the imbalance would be to deny the "o" (subordinate) produced "x" (good job performance) relationship. This example describes a relationship between a white supervisor and a black subordinate. The same propositions would apply if the roles were reversed.

Research on balance theory indicates that structures consisting of balanced relations are perceived as more pleasant than structures consisting of unbalanced relations (Morrisette, 1958; Price, Hamberg & Newcomb, 1966). However there does not appear to be any research on the application of balance theory to performance evaluation. In the present research an attempt is made to relate balance theory to rating behavior.

Purpose of the Present Investigation

The purpose of the present investigation was to assess the effects of race of rater and ratee on favorability of performance ratings when the quality of ratee performance was systematically controlled. The question was whether raters would be more lenient when rating persons of their own race. In previous investigations relevant to this question
(Bass & Turner, 1973; Hamner et al., 1974), the quality of ratee performance was not adequately controlled. This investigation was also aimed at identifying those processes which may underly racial bias in criterion ratings.
CHAPTER II

METHOD

Subjects

The subjects for this investigation were 303 undergraduate students from the University of Maryland, College Park (94 white males, 79 black males, 80 white females, and 50 black females). The college experience of the students is shown in Table 1. This data could only be obtained on 89% of the subjects. Comparisons of the ratings provided by the four groups indicated no significant differences. All of the subjects were volunteers.

Procedure

All of the data were collected prior to or immediately following the students' regularly scheduled classes. The students were enrolled in undergraduate psychology, sociology, and Afro-American studies courses. The data were collected by a black researcher. The primary task of the participants was to review and evaluate the teaching performance of eight hypothetical instructors. Information about each instructor was presented in an 8 X 11 inch manilla envelope. Each envelope contained a page of narrative information along with a 5 X 7 inch photograph.

All of the hypothetical instructors were males between the ages of twenty-two and thirty. Only male instructors were used in order to narrow the focus of the research and facilitate clearer interpretation.
Table 1

College Experience of Subjects

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Freshmen</th>
<th>Sophomores</th>
<th>Juniors</th>
<th>Seniors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black Males</td>
<td>22</td>
<td>15</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>White Males</td>
<td>17</td>
<td>17</td>
<td>37</td>
<td>23</td>
</tr>
<tr>
<td>Black Females</td>
<td>15</td>
<td>16</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>White Females</td>
<td>15</td>
<td>15</td>
<td>33</td>
<td>17</td>
</tr>
</tbody>
</table>
of the results. Five of the eight instructors were white and three were black. This was done in order not to give away the main thrust of the study, which was to compare evaluations of black and white instructors. An equal number of black and white instructors would have appeared highly suspicious since such proportions rarely exist in reality.

The narrative information about each instructor consisted of a list of ten statements describing the past teaching performance. The list of statements was either:

1. Highly favorable, with the combined list having a mean value of 4.4 on a five point scale;
2. Highly unfavorable, with the combined list having a mean value of 1.4 on a five point scale, or
3. Neither highly favorable nor highly unfavorable, with the combined list having a mean value of 3.0 on a five point scale.

Scale values for the statements were computed on the basis of ratings obtained from a sample of 42 undergraduate students. These students filled out the questionnaire shown in Appendix 1. The sample consisted of 22 white and 20 black students. In order to assess the consistency of the ratings, responses from one group were correlated with responses from the other. The average rating assigned to each statement was used for this comparison. The correlation obtained was .97. As an additional check, the two groups were divided and ratings from ten white students were combined with the ratings from eleven black students. A correlation of .98 was obtained. These correlations indicate that ratings of item favorability were very reliable. Utilizing the scaled statements, three levels of teaching performance, good, average, and poor were described. There were two white instructors
and one black instructor in both the good and poor performance categories. The average performance category consisted of one white and one black instructor (Appendix 2).

In addition to manipulating the race and performance level of the instructors, the order of presentation of information was also varied. The design is shown in Table 2. Table 2 shows that approximately half of the subjects were presented a photograph followed by the list of statements. The remainder of the students were presented the list of statements followed by the photograph. The students were randomly assigned to each of these groups. They were required to evaluate the performance of each instructor immediately after reviewing all the information presented about him. The students were given a special rating form for rating the instructors (See Appendix 3). Each rating form was filled out immediately after the student had reviewed information about an instructor. As shown in Appendix 3, the students were asked to respond to the following questions concerning each instructor.

1. Based on the information provided, how well do you think this instructor performs his job?

<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>very poorly</td>
</tr>
<tr>
<td>2</td>
<td>poorly</td>
</tr>
<tr>
<td>3</td>
<td>satisfactorily</td>
</tr>
<tr>
<td>4</td>
<td>well</td>
</tr>
<tr>
<td>5</td>
<td>very well</td>
</tr>
</tbody>
</table>

2. On the basis of the information provided, would you recommend this instructor for promotion?

<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>definitely not</td>
</tr>
<tr>
<td>2</td>
<td>probably not</td>
</tr>
<tr>
<td>3</td>
<td>uncertain</td>
</tr>
<tr>
<td>4</td>
<td>probably</td>
</tr>
<tr>
<td>5</td>
<td>definitely</td>
</tr>
</tbody>
</table>

In addition to evaluating the teaching performance of the instructors, perceived belief similarities between the students and the instructors were also measured. Students were asked the following questions regarding the instructors:
Table 2

Order of Presentation of Information

<table>
<thead>
<tr>
<th>N</th>
<th>Race</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>White</td>
<td>Ph N</td>
</tr>
<tr>
<td>90</td>
<td>White</td>
<td>N Ph</td>
</tr>
<tr>
<td>67</td>
<td>Black</td>
<td>Ph N</td>
</tr>
<tr>
<td>62</td>
<td>Black</td>
<td>N Ph</td>
</tr>
</tbody>
</table>

Ph = Photograph

N = Narrative Information
3. If this instructor were to explain to you his **philosophy** about teaching, how likely do you think it is that you would agree with him?

1 2 3 4 5
would would would would would
definitely probably neither probably definitely
disagree disagree agree nor agree agree
disagree

4. Based upon the information provided about this person, how likely do you think it is that you would agree with his **beliefs** about students?

1 2 3 4 5
would would would would would
definitely probably neither probably definitely
disagree disagree agree nor agree agree
disagree

The questions above were designed to determine the extent to which the students assumed that their beliefs were different from those of instructors of a different race. The students were also asked the following question about each instructor.

5. If this instructor's performance changes in the future, in which direction do you think the change will occur?

1 2 3 4 5
his his I do not his his
performance performance think it performance performance
will decline will decline will change will will
significantly a little improve a little improve
In order to elicit the full cooperation of the students, their classroom instructor was asked to:

(a) inform students about the experiment at least one class period ahead of time
(b) emphasize that participation was voluntary
(c) inform the students that experimental sessions would last approximately 45 minutes
Method of Data Collection

The researcher was introduced to the class at the outset of each experimental session. The researcher then said:

A graduate student here at the university is conducting a study of the processes involved in teacher evaluations. As part of this research, we have been going out to students and having them evaluate the teaching performance of several fictitious faculty members. What we would like to do today is present to you brief descriptions of eight persons all of whom are fictitious. You are to assume however that all eight persons are actual faculty members at a university. On the basis of brief descriptions of these persons, which will be provided to you, we would like for you to evaluate their teaching performance. We will pass out to each of you, eight envelopes. Each envelope will contain two pages of information and a rating form. One of the pages of information will consist of a list of short statements describing the manner in which the person typically conducts himself. The other page consists of a photograph of the person. When the envelopes are given to you, we would like for you to open them one at a time. Once you have opened an envelope you should read the information contained inside, and assuming the person were a faculty member, use the rating form to evaluate his teaching performance. Do you have any questions at this point?

We will pass out the envelopes now. Please do not open any of the envelopes until you are told to do so. When you are told to begin, you should (a) Open one of the envelopes and remove the information inside; (b) Read the information about the person to be evaluated, (c) Once you have read the information evaluate the person utilizing the rating form provided. Remember that you should read the information about the instructor before attempting to evaluate him. It is important that you do not go back over any of the information on a page once you have read it. We would like to get your initial reaction to the information. Once you have filled out the first rating form, put the rating form and the descriptive information back into the envelope. Set that envelope aside and take another. Repeat steps (a) through (c) until you have evaluated all eight persons. When you have finished, we will pass out a short questionnaire for you to fill out. Do you have any questions? You may open the first envelope and begin.

Debriefing

After all of the hypothetical instructors had been evaluated, each of the students was asked to fill out the questionnaire shown in Appendix 4. The purpose of the questionnaire was to determine:
1. whether manipulation of the race of the hypothetical instructors was effective

2. what the students thought about as they rated each instructor

After the questionnaires had been completed and all of the materials collected, the students were told what the research was about. They were told that the study was designed to assess the degree to which they could evaluate each instructor's performance solely on the basis of the narrative descriptions without regard to the photographs. The researcher then answered all questions concerning the study.

Research Hypotheses

Based upon the research of Flaugher et al. (1972) and Hamner et al. (1974), the following hypotheses were made.

Hypothesis I. Independent of performance there will be a significant correlation between the racial similarity of rater and ratee and evaluations of performance. Ratings of black instructors will be higher when rated by blacks and ratings of white instructors will be higher when rated by whites.

Hypothesis II. There will be a significant correlation between the order of presentation of information and ratings of performance. The photographs will act as negative information where the race of rater and ratee are not the same.

The following hypothesis follows from the research of Byrne and Wong (1962), Bixentine and Hawkins (1971), Insko and Robinson (1967), Rokeach and Miezer (1967), Rokeach, Smith and Evans (1960), and Senn (1971).
Hypothesis III. Independent of performance, there will be a significant correlation between the racial combination of rater and ratee and judgements of belief similarity. The subjects will perceive greater belief similarity with hypothetical instructors of the same race than with instructors of a different race.

The following hypothesis follows from the research of Burnstein (1967), Jordan (1953), Morrisette (1958), and Price, Hamberg and Newcomb (1966).

Hypothesis IV. Where the relationship between favorability of information, race of the subject and race of the instructor are balanced, there will be significantly less expected future change in teaching performance. Balanced vs. unbalanced structures will be defined as follows:

1. white subject + white instructor + favorable information = balanced
2. black subject + black instructor + favorable information = balanced
3. white subject + black instructor + unfavorable information = balanced
4. black subject + white instructor + unfavorable information = balanced
5. white subject + black instructor + favorable information = unbalanced
6. black subject + white instructor + favorable information = unbalanced
7. white subject + white instructor + unfavorable information = unbalanced
8. black subject + black instructor + unfavorable information  
    = unbalanced
CHAPTER III

RESULTS

Primary Findings and Post Hoc Analyses

Manipulation of the Independent Variables

This study involved the manipulation of four independent variables (a) the race of the hypothetical instructor, (b) the level of the instructor’s performance, (c) the order in which the information was presented and (d) race of the rater.

Race of the hypothetical instructor. Among the eight hypothetical instructors, five were white and three were black. Manipulation of race was accomplished through the use of photographs. The effect of the manipulation was measured by the students' responses to the following question: "How many of the instructors would you guess could be classified as members of a minority group?" This question was asked after all of the hypothetical instructors had been evaluated and the material describing them returned to the original envelope. Thirty-seven percent of the students remembered the exact number of black instructors. Fifty-three percent of the students indicated the correct number plus or minus one (either two or four). The remainder of the students indicated that they could not remember the number of black instructors rated.

The level of the instructor's performance. Table 3 shows the mean performance ratings given to instructors who were described as good, average or poor performers. The ratings were based on a five point
Table 3

Mean Performance Ratings of Good, Average and Poor Instructors

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Race of Instructor</th>
<th>Mean Performance Rating</th>
<th>Quality of Performance Promotability</th>
</tr>
</thead>
<tbody>
<tr>
<td>High*</td>
<td>White</td>
<td>4.48</td>
<td>4.28</td>
</tr>
<tr>
<td>High</td>
<td>Black</td>
<td>4.50</td>
<td>4.36</td>
</tr>
<tr>
<td>Average</td>
<td>White</td>
<td>2.85</td>
<td>2.69</td>
</tr>
<tr>
<td>Average</td>
<td>Black</td>
<td>2.93</td>
<td>2.98</td>
</tr>
<tr>
<td>Low*</td>
<td>White</td>
<td>1.47</td>
<td>1.46</td>
</tr>
<tr>
<td>Low</td>
<td>Black</td>
<td>1.38</td>
<td>1.44</td>
</tr>
</tbody>
</table>

*The average rating of the two white instructors.
scale (questions one and two, Appendix 3). It can be seen from Table 3 that the students clearly differentiated between good, average, and poor instructors.

Order of presentation of information. Order effects were hypothesized on the basis of the assumptions that (a) photographs of black instructors represented negative information to white raters and positive information to black raters, and (b) photographs of white instructors represented negative information to black raters and positive information to white raters. Combining the photographs with highly favorable or highly unfavorable narrative information was expected to produce positive-positive, positive-negative, negative-positive, or negative-negative order effects. Observation of the students as they reviewed the material revealed that manipulation of the order of presentation of information was not very effective. Many students ignored the order in which the material was presented, and looked at the photograph prior to reading the narrative information.

Race of rater. Subjects were classified as white or black solely on the basis of their physical characteristics. At no time were the subjects asked to provide information concerning their race. However, as it turned out, the subjects were enrolled in classes that were overwhelmingly white or overwhelmingly black.

Assessment of the Interdependence of Ratings

There were 303 subjects, each of whom rated (the same) eight hypothetical instructors. Ratings of the two white instructors in the high performance category were combined and the average rating used for analysis. This was also done for the two white instructors in the poor performance category. As a result, there were six observations per
subject instead of eight. In computing the correlations discussed in this chapter, each rating was treated as an independent measure. The fact that the same subject produced more than one rating was ignored. As an example, when racial similarity between rater and ratee was correlated with ratings of teaching performance, (using the total sample) an N of 1818 (303 X 6) was used. This was done only after determining that there were no significant differences between subjects with regard to rating style. An analysis of variance with repeated measurements was used to determine whether there was a significant difference between subjects with regard to the type of ratings given. The design is shown in Figure 1. The analysis was done separately on the ratings of white males, white females, black males and black females. Table 4 shows the results of white males' ratings of white instructors. The dependent variable was how well the students thought each instructor performed his job.

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_3$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_4$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_n$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Level of performance = high, medium, and low

Figure 1. Analysis of variance repeated measurements design.
Table 4

Summary Table of the Performance Ratings of White Instructors by White Males

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>282</td>
<td>556</td>
<td>.28</td>
<td>.82</td>
</tr>
<tr>
<td>S</td>
<td>94</td>
<td>26.4</td>
<td>.314</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>465.98</td>
<td>233</td>
<td>685*</td>
</tr>
<tr>
<td>S/A</td>
<td>186</td>
<td>63.6</td>
<td>.34</td>
<td></td>
</tr>
</tbody>
</table>

*p < .01
As Table 4 shows, performance level accounted for all of the significant variance in performance ratings. There were no significant subject effects. Table 5 shows the results of black males' ratings of black instructors in terms of quality of performance.

As Table 5 shows, performance level again accounted for all of the significant variance in performance ratings. There were no significant subject effects. The results were similar for white females and black females. Performance level accounted for all of the significant variance in performance ratings. Again there were no subject effects.

All of the results are reported on the basis of the race of the rater. None of the results are reported by sex of rater. That is, white raters include both males and females. Black raters also include both males and females. This was done after comparing male and female raters. Zero order correlations were computed (by sex of the rater) between rater-ratee racial similarity and (a) ratings of performance and, (b) ratings of belief similarity. The correlations for female raters were compared with those of male raters. None of the correlations was significantly different.

Order of Presentation of Information and Performance Ratings

Partial correlations were computed between order of presentation and performance ratings, with level of performance and race of the ratee held constant. The correlations were computed on the basis of the total sample and separately for black and white raters. None of the correlations was significant.

Level of Performance and Performance Ratings

The correlations presented in Table 6 represent the relationship between level of performance (low, average, and high) and performance
Table 5

Summary Table of the Performance Ratings of Black Instructors
By Black Males

<table>
<thead>
<tr>
<th>SV</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>233</td>
<td>419</td>
<td>.65</td>
<td>.76</td>
</tr>
<tr>
<td>S</td>
<td>77</td>
<td>49.67</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>239.14</td>
<td>119.57</td>
<td>140.67*</td>
</tr>
<tr>
<td>S/A</td>
<td>154</td>
<td>130.9</td>
<td>.85</td>
<td></td>
</tr>
</tbody>
</table>

*p < .01
Table 6

Zero Order Correlations Between Level of Performance of Performance Ratings

<table>
<thead>
<tr>
<th>Race of Rater</th>
<th>Quality of Performance</th>
<th>Promotability</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>.92**</td>
<td>.88*</td>
</tr>
<tr>
<td></td>
<td>(1037)</td>
<td>(1037)</td>
</tr>
<tr>
<td>Black</td>
<td>.85**</td>
<td>.82**</td>
</tr>
<tr>
<td></td>
<td>(757)</td>
<td>(756)</td>
</tr>
<tr>
<td>Total</td>
<td>.89**</td>
<td>.85**</td>
</tr>
<tr>
<td></td>
<td>(1794)</td>
<td>(1793)</td>
</tr>
</tbody>
</table>

***p < .01

Note: Sample size shown in parentheses
ratings. Level of performance accounted for a large proportion of the variance (67% - 85%) in both the black and white groups. Since the major research hypotheses of the present study are concerned with contaminants beyond these relevant aspects of performance, level of performance will be partialled out of the ratings before testing the hypotheses.

The Relationship Between Rater-Ratee Similarity and Performance Ratings With Other Variables Held Constant

Partial correlations were computed between ratings of performance (questions one and two Appendix 3) and racial similarity between rater and ratee, with order of presentation of information and level of performance held constant. These correlations were computed for both white and black students, and the total sample. Partial correlations based upon the total sample represent the degree to which racial similarity between rater and ratee was related to performance ratings without regard to the specific race of rater or ratee. Dummy coding (0 = dissimilar, 1 = similar) was used to designate rater-ratee similarity. The dependent variable was based on the five point scales shown in Appendix 3. A partial correlation of .07 was obtained between rater-ratee similarity and ratings of how well each instructor performed his job. This correlation is significant at the .05 level with 1790 degrees of freedom. The partial correlation between rater/ratee similarity and ratings of promotability was .04, df = 1789, p < .05. The partial correlations broken down by race of rater are shown in Table 7. These correlations actually represent the degree of relationship between the instructor's race and his performance rating, since the race of the rater was held constant.
Table 7

Second Order Partial Correlations Between Race of Ratee and Performance Ratings by Race of Rater

<table>
<thead>
<tr>
<th>Race of Rater</th>
<th>Quality of Performance</th>
<th>Promotability</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>.08*</td>
<td>-.02</td>
<td>1037</td>
</tr>
<tr>
<td>Black</td>
<td>.06</td>
<td>.12**</td>
<td>755</td>
</tr>
</tbody>
</table>

*p < .05
**p < .01
White students gave higher ratings to instructors of the same race when evaluating the quality of the instructors' performance. Black students gave higher ratings to instructors of the same race when assessing the instructors' promotability.

Racial Similarity of Rater and Ratee Correlated with Ratings of Belief Similarity.

Based upon the total sample, the partial correlation between rater-ratee racial similarity and agreement with the instructor's teaching philosophy was .05 (df = 1788, p < .01) with the order of presentation of information and level of performance held constant. The partial correlation between racial similarity and agreement with the instructors beliefs about students was .06 (df = 1788, p < .01). The same correlations broken down by race of the rater are shown in Table 8.

Table 8 shows that the race of ratee was significantly related to perceived agreement with the instructor's beliefs in three out of four cases. In all three cases there was greater agreement with the beliefs of instructors of the same race as the rater.

Expected Change in Future Performance

In addition to evaluating teaching performance, the students provided ratings of how much they expected each instructor's performance to change. The responses were analysed by computing the average rating given to each instructor by white and black students. Comparisons were made between balanced and unbalanced situations as defined on page 28. The results are shown in Table 9. There were no significant differences between the ratings based upon balanced-unbalanced categorization.
Table 8

Second Order Partial Correlations Between Race of the Instructor and Ratings of Belief Similarity by Race of Rater

<table>
<thead>
<tr>
<th>Race of Rater</th>
<th>Philosophy About Teaching</th>
<th>Beliefs About Students</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>.07**</td>
<td>.05</td>
<td>1037</td>
</tr>
<tr>
<td>Black</td>
<td>.08*</td>
<td>.09**</td>
<td>755</td>
</tr>
</tbody>
</table>

*p < .05

**p < .01
Table 9

Average Rating of Expected Change by Race and Degree of Cognitive Balance

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Race of Rater</th>
<th>Race of Ratee</th>
<th>Cognitive Definition</th>
<th>$\bar{X}$ Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>White</td>
<td>White</td>
<td>Balanced</td>
<td>3.74</td>
</tr>
<tr>
<td>High</td>
<td>Black</td>
<td>White</td>
<td>Unbalanced</td>
<td>3.83</td>
</tr>
<tr>
<td>High</td>
<td>Black</td>
<td>Black</td>
<td>Balanced</td>
<td>4.04</td>
</tr>
<tr>
<td>High</td>
<td>White</td>
<td>Black</td>
<td>Unbalanced</td>
<td>3.87</td>
</tr>
<tr>
<td>Low</td>
<td>White</td>
<td>Black</td>
<td>Balanced</td>
<td>2.69</td>
</tr>
<tr>
<td>Low</td>
<td>Black</td>
<td>Black</td>
<td>Unbalanced</td>
<td>2.78</td>
</tr>
<tr>
<td>Low</td>
<td>Black</td>
<td>White</td>
<td>Balanced</td>
<td>2.70</td>
</tr>
<tr>
<td>Low</td>
<td>White</td>
<td>White</td>
<td>Unbalanced</td>
<td>2.82</td>
</tr>
</tbody>
</table>
Post Hoc Analyses

Having found that there was a significant relationship between rater-ratee racial similarity and performance ratings, further analyses were conducted to determine the performance level at which the relationship was strongest. No hypotheses had been made with respect to these analyses.

Partial correlations were computed between rater-ratee racial similarity and performance ratings for each level of teaching performance, with the order of presentation of information held constant. The results are shown in Table 10. White raters gave higher ratings (in terms of promotability) to average black instructors than to average white instructors. They rated poor white instructors more favorably than poor black instructors in terms of quality of performance. Black raters evaluated average black instructors more favorably than average white instructors both with respect to quality of performance and promotability. Table 11 shows the mean ratings for each level of teaching performance by race of rater and ratee.

Table 12 shows the partial correlation between race of the instructor and ratings of belief similarity for each level of teaching performance. The data indicates that racial similarity of rater and ratee was related to perceived belief similarity only within the average category for white raters and the poor category for black raters. White raters indicated greater agreement with the beliefs of a mediocre white instructor than a similar black instructor. Black raters showed greater agreement with the beliefs of an average black instructor than a similar white instructor. Table 13 shows the mean ratings of belief similarity for each level of teaching performance by race of rater and ratee.
Table 10
Partial Correlations Between Instructor’s Race and Performance Ratings by Level of Performance

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Type of Rating</th>
<th>Whites (N = 348)</th>
<th>Blacks (N = 249)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Quality of Performance</td>
<td>.01</td>
<td>.05</td>
</tr>
<tr>
<td>High</td>
<td>Promotability</td>
<td>-.03</td>
<td>.08</td>
</tr>
<tr>
<td>Average</td>
<td>Quality of Performance</td>
<td>.04</td>
<td>.13*</td>
</tr>
<tr>
<td></td>
<td>Promotability</td>
<td>-.11*</td>
<td>.22**</td>
</tr>
<tr>
<td>Low</td>
<td>Quality of Performance</td>
<td>.17**</td>
<td>-.02</td>
</tr>
<tr>
<td></td>
<td>Promotability</td>
<td>.08</td>
<td>.03</td>
</tr>
</tbody>
</table>

*p < .05
**p < .01
### Table II

Mean Ratings of Performance by Race of Rater and Ratee

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Type of Rating</th>
<th>Race of Ratee</th>
<th>White Raters</th>
<th>Black Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean (X)</td>
<td>Standard Deviation (SD)</td>
</tr>
<tr>
<td>High</td>
<td>Quality of Performance</td>
<td>White</td>
<td>4.58</td>
<td>.461</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Black</td>
<td>4.56</td>
<td>.601</td>
</tr>
<tr>
<td></td>
<td>Promotability</td>
<td>White</td>
<td>4.28</td>
<td>.548</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Black</td>
<td>4.32</td>
<td>.662</td>
</tr>
<tr>
<td>Average</td>
<td>Quality of Performance</td>
<td>White</td>
<td>2.93</td>
<td>.673</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Black</td>
<td>2.87</td>
<td>.723</td>
</tr>
<tr>
<td></td>
<td>Promotability</td>
<td>White</td>
<td>2.76</td>
<td>.754</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Black</td>
<td>2.96</td>
<td>.773</td>
</tr>
<tr>
<td>Low</td>
<td>Quality of Performance</td>
<td>White</td>
<td>1.38</td>
<td>.535</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Black</td>
<td>1.21</td>
<td>.454</td>
</tr>
<tr>
<td></td>
<td>Promotability</td>
<td>White</td>
<td>1.40</td>
<td>.603</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Black</td>
<td>1.31</td>
<td>.575</td>
</tr>
</tbody>
</table>

* Means are significantly different at the .05 level.
Table 12

Partial Correlations Between Race of Ratee and Ratings of Belief Similarity for Each Level of Teaching Performance

<table>
<thead>
<tr>
<th>Level of Performance</th>
<th>Type of Rating</th>
<th>Whites (N = 348)</th>
<th>Blacks (N = 249)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Agreement with Instructors' Philosophy about Teaching</td>
<td>-.05</td>
<td>.09</td>
</tr>
<tr>
<td>High</td>
<td>Agreement with Instructors' Philosophy about Students</td>
<td>.04</td>
<td>.06</td>
</tr>
<tr>
<td></td>
<td>Agreement with Instructors' Philosophy about Teaching</td>
<td>.01</td>
<td>.18*</td>
</tr>
<tr>
<td>Average</td>
<td>Agreement with Instructors' Philosophy about Students</td>
<td>.03</td>
<td>.22*</td>
</tr>
<tr>
<td></td>
<td>Agreement with Instructors' Philosophy about Teaching</td>
<td>.12*</td>
<td>-.05</td>
</tr>
<tr>
<td>Low</td>
<td>Agreement with Instructors' Philosophy about Students</td>
<td>.15*</td>
<td>-.06</td>
</tr>
</tbody>
</table>

*p < .01
Table 13

Mean Ratings of Belief Similarity by Race of Rater and Ratee

| Performance | Type of Rating | Race of Ratee | White Raters | | Black Raters | |
|-------------|----------------|---------------|--------------| |--------------| |
|             | Teaching       | White         | 4.10         | .521         | 4.04         | .612         |
|             |                 | Black         | 4.16         | .590         | 4.16         | .732         |
| High        | Philosophy     | White         | 4.12         | .570         | 4.03         | .667         |
|             |                 | Black         | 4.17         | .660         | 4.11         | .719         |
|             | Beliefs About   | White         | 2.81         | .816         | 2.66         | .865         |
|             | Students       | Black         | 2.79         | .834         | 3.00*        | .930         |
| Average     | Teaching       | White         | 3.04         | .870         | 2.62         | .925         |
|             | Philosophy     | Black         | 2.99         | .919         | 3.04*        | .968         |
|             | Beliefs About   | White         | 1.66         | .657         | 1.77         | .736         |
|             | Students       | Black         | 1.50*        | .677         | 1.70         | .897         |
| Low         | Teaching       | White         | 1.77         | .720         | 1.74         | .723         |
|             | Philosophy     | Black         | 1.56*        | .666         | 1.67         | .624         |

*Means are significantly different at the .05 level.
Process Variables

The questionnaires completed after all of the instructors had been rated were designed to measure those process variables which may have influenced the students' ratings. Questions 1, 7 and 11, (Appendix 4), were specifically designed for this purpose.

Question 7 was designed to determine whether there were differences between black and white students in terms of their criteria for evaluating teaching performance. The question was, "What do you think is the most important thing for an instructor to accomplish in teaching a course". Content analysis of the responses yielded the results shown in Table 14. As Table 14 shows, both groups felt that the primary goal of an instructor should be to teach in a manner which helps students understand the material. Black students also felt that an instructor should attempt to develop a good relationship with students, by trying to understand them and show consideration for their needs. White students felt it was important for an instructor to present the material in an interesting manner so that students want to learn. In general, white students provided a greater variety of responses than black students.

Table 15 shows students' responses to the question, "As you were actually rating the instructors, what thoughts went through your mind". Table 15 shows that most of the students thought about the consequences of being taught by each instructor as they evaluated him. The primary concern seemed to be the consequences each instructor's performance would have on his students. Thoughts along this line were apparently crystallized by associating the individual being rated with past instructors. A fair percentage of the students wondered about the real purpose of the experiment.
Table 14

What Students Thought Most Important for an Instructor to Accomplish

<table>
<thead>
<tr>
<th>Response</th>
<th>Blacks</th>
<th></th>
<th></th>
<th>Whites</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>% of Total Responses</td>
<td>Number</td>
<td>% of Total Responses</td>
<td></td>
</tr>
<tr>
<td>1. Present the material in a manner which facilitates understanding</td>
<td>62</td>
<td>59</td>
<td>99</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>2. Develop a rapport with students (understanding, respect, consideration)</td>
<td>17</td>
<td>16</td>
<td>26</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>3. Present the material in an interesting manner, motivate the student</td>
<td>7</td>
<td>7</td>
<td>59</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>4. Treat students fairly in grading etc., show no favoritism</td>
<td>7</td>
<td>7</td>
<td>14</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>5. Demonstrate the relevance of the material</td>
<td>5</td>
<td>5</td>
<td>19</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>6. Set course objectives, let students know what is expected of them</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>7. Relate the lecture to the tests, follow course outline, provide resource material</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8. Keep good office hours</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>9. Have knowledge of the subject matter</td>
<td>-</td>
<td>-</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10. Get feedback from students</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Response</td>
<td>Blacks</td>
<td></td>
<td></td>
<td>Whites</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>--------</td>
<td>----------</td>
<td>----------</td>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td>Number</td>
<td>% of Total Responses</td>
<td>Number</td>
<td>% of Total Responses</td>
<td></td>
</tr>
<tr>
<td>1. How I would feel if I were in the instructor's class. How I felt when I had instructors similar to these</td>
<td>47</td>
<td>64</td>
<td>92</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>2. What was the research really about</td>
<td>7</td>
<td>10</td>
<td>24</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>3. The way the instructor was described</td>
<td>7</td>
<td>10</td>
<td>27</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>4. The fact that there was not enough information about each instructor</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5. Whether or not I was being fair in my rating</td>
<td>2</td>
<td>2</td>
<td>11</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>6. Whether the picture matched the descriptions</td>
<td>2</td>
<td>2</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>7. Why there were no female instructors</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8. How time consuming the task was</td>
<td>2</td>
<td>2</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>9. Whether or not the person met my standards</td>
<td>2</td>
<td>2</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>10. How other students would rate the Black instructors</td>
<td>2</td>
<td>2</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>11. How it felt to be able to judge an instructor instead of vice versa</td>
<td>--</td>
<td>--</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12. Whether younger instructors were described more favorably</td>
<td>--</td>
<td>--</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>13. Whether or not the instructor will change</td>
<td>--</td>
<td>--</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Question one was designed to measure the students' "guesses" as to what the research was actually about. Approximately 19 percent of the white students and 5 percent of the black students thought the research "had something to do with" comparing the ratings of white and black instructors.

Summary of Data Supportive and Non-Supportive of the Research Hypotheses

Hypothesis I. Independent of the level of performance there will be a significant correlation between the racial similarity of rater and ratee and evaluations of performance. Ratings of black instructors will be higher when rated by blacks and ratings of white instructors will be higher when rated by whites.

Hypothesis I was supported by the results from this study. Evaluations of teaching performance were related to the racial similarity of rater and ratee. However, racial similarity of rater and ratee accounted for less than one percent of the variance in ratings of teaching quality or promotability.

Hypothesis II. There will be a significant correlation between the order of presentation of information and ratings of performance. The photographs will act as negative information where the race of the instructor and the race of the rater are not the same. Hypothesis II was not supported by the data obtained from this study.

Hypothesis III. Independent of the level of performance there will be a significant correlation between the racial similarity of rater and ratee and judgments of belief similarity. Subjects will perceive greater belief similarity with hypothetical instructors of the same race than with other instructors.
Hypothesis III was supported by the data from this research. The degree to which students thought their beliefs and those of an instructor were similar was related to the instructor's race. However, racial similarity of rater and ratee accounted for less than one percent of the variance in ratings of agreement with teaching philosophy or beliefs about students.

Hypothesis IV. Where the relationship between favorability of information, race of the subject and the instructor's race is balanced, there will be significantly less expected future change in the instructor's teaching performance. Manipulation of the independent variable was not sufficient to test this hypothesis. That is, the subjects did not seem to form cognitive triads on the basis of each instructor's race. Responses to the debriefing questionnaire indicated the consequences each instructor's teaching performance would have on students was the most salient factor influencing the formation of cognitions about him.
CHAPTER IV

DISCUSSION

Before discussing the results of this research, several methodological issues should be addressed. First, the effects of the race of the researcher are not known. Since only a black researcher was present at the time the data was collected, experimenter effects may have been operative. The subjects' rating behavior may have been different had a team of researchers (one white and one black) collected the data. It is also important to note that the levels of teaching performance were for the most part extreme. Descriptions of good instructors left little doubt as to the quality of their teaching performance as did descriptions of poor instructors. Since such wide variations in the performance ratings were experimentally induced, there may have been little opportunity for other effects to have a major impact. A third methodological consideration is that the subjects were not selected at random. The participants were volunteers from existing college classes. The degree to which their rating behavior was typical of college students in general is not known. Moreover, the generalizability of the results of this research to non-educational organizations is also unknown. Taken singly, or in total, the factors outlined above might have influenced the outcomes of this research. In pointing out these factors, the intent is not to detract from the validity of the findings, but rather to facilitate appropriate interpretation of the data.
The results of this study indicate that teaching performance was the most important variable affecting rating behavior. When the entire range of teaching performance is considered, race effects were small. This could have been expected to some degree, since descriptions of teaching performance in the good and poor categories were extreme. Nevertheless, on the basis of the total sample, level of performance accounted for 79% of the variance in ratings of teaching quality and 72% of the variance in ratings of promotability. These findings are corroborated by the fact that less than 50% of the subjects indicated remembering the exact number of black instructors rated. The data do not however support the conclusion that there were no race effects.

Hypothesis I was supported by the results in Table 7. Racial similarity of rater and ratee was significantly related to performance ratings for both white and black subjects when the level of teaching performance was statistically controlled. Higher ratings were given when the rater and ratee belonged to the same race. This finding is supported by the research of Flaugher et al. (1969). The race effects shown in Table 7 are so small however that use of controls such as forced choice rating scales or preliminary training for the raters might have completely eliminated rater bias altogether. The basic finding for both white and black subjects was that when given relevant information, the subjects could accurately evaluate performance. The effects of irrelevant information such as the race of the ratee were minimal.

Hypothesis II was not sufficiently tested by this study. It is not surprising that order of presentation of stimulus information was insignificant, since observation of the raters revealed that manipulation of this variable was not effective. Many students ignored the order in which the material was presented and looked at the photographs prior to
evaluating each instructor. In future research it would be better to have the information about each ratee presented on film or slides thus allowing greater control over order of presentation.

Hypothesis III is supported by the results in Table 8. Ratings of belief similarity were related to the racial similarity of rater and ratee for both black and white subjects when level of teaching performance was statistically controlled. Raters assumed that their beliefs were more similar to those of an instructor of the same race than an instructor of a different race. However, ratings of belief similarity closely paralleled ratings of teaching performance and may simply represent the halo effect. Despite this possibility, ratings of belief similarity are especially interesting since no information was given about the beliefs of any of the instructors. The subjects apparently made assumptions about each instructor's beliefs without relevant information. The assumptions were based to some extent on each instructor's race. This finding is supported by the research of Byrne and Wong (1971), Inako and Robinson (1967), Rokeach and Miezer (1967), Rokeach, Smith and Evans (1960), and Senn (1971).

Hypothesis IV was not adequately tested by this research. The degree of cognitive balance, as defined in this study, was not sufficiently manipulated to determine the applicability of balance theory to rater bias. This was probably due to the fact that in forming cognitions about each instructor, the students gave more weight to the consequences of an instructor's behavior than to his race. The students apparently did not experience positive or negative sentiments toward an instructor solely on the basis of his race. This was the assumption underlying the definitions of balance employed here. The primary concern of the students seemed to be the consequences each instructor's behavior would
have for students. Black raters, for example, did not necessarily perceive as unbalanced, situations where a black instructor was described unfavorably or a white instructor described favorably. In addition, white raters did not necessarily perceive as unbalanced, situations where a white instructor was described unfavorably or a black instructor described favorably.

**Implications for Future Research**

Post hoc analysis of the data yielded some useful information. Apparently there were no significant race effects where teaching performance was described favorably; but where teaching performance was described as average, white raters gave higher ratings to black instructors than to white instructors in terms of promotability. White raters also gave poor white instructors higher ratings than poor black instructors in terms of quality of performance. Black raters gave higher ratings to average black instructors than to average white instructors both in terms of quality of performance and promotability. Apparently white raters were influenced by the race of the ratee only where teaching performance was not clearly outstanding. The magnitude and direction of the relationship seems to have depended on a number of factors including the rating situation and the performance dimension to be evaluated. White raters may have been more willing to promote an average black instructor than an average white instructor for any one or all of the following reasons: (a) This performance dimension was more subjective than the others; (b) These raters were more willing to give a black instructor the benefit of the doubt; (c) Given equivalent teaching ability, the raters felt it more important from an affirmative action standpoint to promote a black instructor; (d) The raters were over-zealous
in their attempts to be fair; (e) As long as teaching performance was not too poor, the raters attempted to rate black instructors as high or higher than white instructors in order to demonstrate that they were not prejudiced. Any of these factors may have accounted for the white raters' evaluations of average instructors. All of them are interesting topics for future research.

The factors underlying white raters' evaluations of poor instructors are also open to speculation. They may have rated poor white instructors more favorably than poor black instructors because they were less willing to accept extremely negative information about individuals of their own race. That is, extremely negative descriptions were probably more in line with white raters' expectations about black instructors. The validity of this hypothesis can only be determined through future research.

Black raters apparently were not influenced by the race of the instructor when rating very good or very poor instructors. However, when evaluating average performance, black raters favored black instructors over white instructors both in terms of teaching performance and promotability. The rationale for their behavior cannot be determined on the basis of the data. They may have felt that an average black instructor deserved higher ratings than an average white instructor by virtue of the obstacles he must face in teaching at a predominantly white university.

In summary then, the degree of rater bias for both white and black subjects seems to have been related to the overall quality of ratee performance. This finding is supported by the research of Flaugher et al. (1969), Hemner et al. (1974), Bigoness (1976), and Rotter and Rotter (reported in Bigoness, 1976). Rotter and Rotter found that low performing
blacks were rated significantly higher than low performing whites, while high performing blacks and whites were rated similarly. Bigoness (1976) found that while low performing males and females were rated similarly, high performing females were rated significantly higher than high performing males. Deaux and Taynor (1973) found that highly competent males were rated more favorably than highly competent females; but males of low competence were rated lower than similar females.

Conclusions

The results of this study indicate that racial bias influenced performance ratings, but only to a limited degree. The subjects in this study evaluated teaching behavior primarily on the basis of relevant information about teaching performance. Teaching performance also influenced ratings of perceived attitude similarity. Other research has shown that the sex of the rater and ratee may influence rating behavior. However, more research must be done before a parsimonious explanation of the effects of rater-ratee characteristics on performance evaluation can be offered. It is clear that in future research greater efforts must be made to assess not only the presence or absence of rater bias, but also those factors which may account for it.
APPENDIX 1

TEACHER EVALUATION QUESTIONNAIRE
As part of my doctoral dissertation research, I am conducting a study involving an evaluation of the teaching performance of faculty members. As part of this project I would like to see how good or bad you think some of the things are that instructors do. Below is a list of statements describing the types of things that instructors do. Read each statement and indicate how good or bad you think the behavior described in the statement is in terms of contributing to effective teaching performance. Please use the following five point scale to rate each item.

1 = Very bad, something an instructor should definitely not do.
2 = Poor, something an instructor should try to avoid doing.
3 = Does not matter, something that really does not matter one way or the other.
4 = Good, something an instructor should try to do whenever possible.
5 = Excellent, something an instructor should definitely do.

1. students usually cannot see as he writes on the board
2. lectures in a low monotone
3. is ill informed about the topics in the course
4. asks for more details when he does not understand what a student says
5. invites students to make suggestions about the course
6. introduces many ideas during each class
7. starts class on time
8. gives concise explanations
9. is enthusiastic
10. invites students to ask questions
11. tells students which material they will not be tested on
12. provides students with the opportunity to apply newly acquired knowledge.
13. synthesizes or integrates the material effectively
14. gives students written comments about their performance
15. gives tests which do not give students the opportunity to show what they have learned
16. does not give enough exams
17. expresses his appreciation when a student does a good job
18. tells the students what parts of the material are the most important
19. gives too many exams
20. addresses students by name
21. can tell if the class is understanding him or not
22. is usually incoherent or vague in what he is saying
23. tells students at the beginning of class what he is going to talk about
24. explains how the topics in the course are related to each other
25. makes class time seem to pass quickly
26. offers topics which are usually relevant to students' educational goals
27. has a well developed plan for each class session
28. gives clear explanations
29. is impartial in grading students
30. makes students feel at ease when talking with him
31. post office hours
32. answers thoroughly all questions which arise during the lecture
33. uses a variety of teaching techniques
34. seldom indicates when a new topic is being introduced
35. is harshly critical of student responses
36. uses understandable vocabulary
37. his lectures follow a logical sequence
38. provides assistance outside of class
39. has favorite students with whom he spends more time than others
40. lectures in such a way that the textbook is easier to understand
41. presents the origins of ideas and/or concepts
42. his lectures make note-taking easy
43. presents examples of what he wants students to do
44. class presentations are well prepared
45. has no interest in the quality of his teaching
46. does not utilize the blackboard very well
47. is tolerant of students' mistakes
48. presents an outline of the course
49. has only a superficial knowledge of the subject matter
50. discusses examination questions and answers before the exams
51. tells students what must be done to get a good grade in the course
52. is not responsive to the students' comments and questions in class
53. does not encourage silent students to participate in class
54. gives exam questions that are very picky
55. uses examples which aid learning of the course material
56. tries to force his own ideas on the student
57. summarizes the material in a way that aids retention
58. fails to differentiate between significant and insignificant materials
59. shows favoritism (has favorites) when he is grading
60. frequently subjects students to ridicule and sarcasm
61. is frequently distracted from the main point while lecturing
62. finds ways to help students answer their own questions
63. discusses the activities persons in his profession engage in
64. explains the grading system
65. is punctual
66. attempts to cover too much material
67. adjusts his pace to the needs of the class
68. does not weight tests, quizzes and reports fairly in determining grades
69. invites students to share their knowledge and experiences
70. the level of the lectures is too difficult
71. does not organize course objectives
72. asks confusing questions on his tests
73. introduces stimulating ideas about the subject
74. admits when he has made a mistake
75. grades tests too strictly
76. seems to lack self confidence
77. does not return tests within a reasonable time.
78. treats all students equally
79. does not grade his tests fairly
80. covers the material in the course too quickly
81. presents recent developments in the field
82. ends class on time
83. rephrases statements the students do not understand
84. asks silent students to participate
85. knows the subject matter
86. does not offer students freedom to express their opinions in class
87. admits when he does not know the answer to a question
88. does not show tolerance for different points of view
89. states when a new topic is being introduced
90. changes his approach to meet new situations
91. never discusses test questions after the tests have been returned
presents up-to-date information

allows students to express their problems related to the course

varies the speed and tone of his voice when lecturing

wastes a lot of time in class

uses students questions as an opportunity for class discussion

invites criticism of the ideas he presents

tends to discourage questions in class

spends too much time in class dwelling on the obvious

gives exams which do not have a balanced coverage of the major topics

lectures in a rambling fashion

does not discuss exam questions after exams have been returned

is usually disrespectful towards students

introduces topics by telling how they relate to topics that have already been discussed

provides informative appraisals of students' work

does not grade tests fairly

talks to the blackboard when lecturing

is not receptive to students' viewpoints

tries to force his own ideas on the students

is harshly critical of student responses

seems to lack self confidence

has only superficial knowledge of the subject matter

gives exams which do not have a balanced coverage of the major topics

answers questions directly and to the point

the way he covers the material in class makes it boring
APPENDIX 2

THE EIGHT HYPOTHETICAL INSTRUCTORS
A Very Good Instructor

He gives clear explanations.

He rephrases statements the students do not understand.

He provides assistance outside of class.

He synthesizes or integrates the material effectively.

He can tell if the class is understanding him or not.

He gives class presentations which are well prepared.

He invites students to make suggestions about the course.

He invites students to ask questions.

He is punctual.

He presents recent developments in the field.
A Very Good Instructor

He knows the subject matter.

He presents up-to-date information.

He uses examples which aid learning of the course material.

He presents examples of what he wants students to do.

He is tolerant of students' mistakes.

He treats all students equally.

He presents an outline of the course.

His lectures make note-taking easy.

His lectures follow a logical sequence.

He is enthusiastic.
A Very Good Instructor

He explains the grading system.

He allows students to express their problems related to the course.

He has a well developed plan for each class session.

He gives concise explanations.

He adjusts his pace to the needs of the class.

He ends class on time.

He tells the students which parts of the material are most important.

He posts office hours.

He explains how the topics in the course are related to each other.

He makes students feel at ease when talking with him.
An Average Instructor

He tells students what has to be done to get a good grade in the course.

He gives students written comments about their performance.

He does not utilize the blackboard very well.

He does not give enough exams.

He is frequently distracted from the main point while lecturing.

He answers thoroughly all questions which arise during the lecture.

He provides students with the opportunity to apply newly acquired knowledge.

He uses a variety of teaching techniques.

He does not encourage silent students to participate in class.

He tries to force his own ideas on the students.
An Average Instructor

He expresses his appreciation when a student does a good job.

He starts class on time.

He asks silent students to participate.

He attempts to cover too much material.

He gives exam questions that are very picky.

He admits when he has made a mistake.

He uses understandable vocabulary.

The level of his lectures is too difficult.

He gives too many exams.

He grades tests too strictly.
A Poor Instructor

He covers the material in the course too quickly.

He does not show tolerance for different points of view.

He does not weight tests, quizzes and reports fairly in determining grades.

He has only superficial knowledge of the subject matter.

He is usually disrespectful towards students.

He seldom indicates when a new topic is being introduced.

He is not receptive to students' viewpoints.

The way he covers the material in class makes it boring.

He asks confusing questions on his tests.

He does not grade his tests fairly.
A Poor Instructor

He does not offer students freedom to express their opinions in class.

He does not organize course objectives.

He is ill-informed about the topics in the course.

He is usually incoherent or vague in what he is saying.

He is harshly critical of student responses.

He fails to differentiate between significant and insignificant material.

Students usually cannot see what he writes on the board.

He does not return tests within a reasonable time.

He has no interest in the quality of his teaching.

He shows favoritism (has favorites) when he is grading.
A Poor Instructor

He has favorite students with whom he spends more time than others.

He wastes a lot of time in class.

He gives tests which do not give students the opportunity to show what they have learned.

He lectures in a rambling fashion.

He frequently subjects students to ridicule and sarcasm.

He talks to the blackboard when lecturing.

He gives exams which do not have a balanced coverage of the major topics.

He lectures in a low monotone.

He does not discuss exam questions after exams have been returned.

He is not responsive to students’ comments and questions in class.
APPENDIX 3

PERFORMANCE EVALUATION RATING FORM
On the basis of the information given to you about this person, please answer the following questions.

1. Based on the information provided, how well do you think this instructor performs his job? (circle one)

   1  2  3  4  5  
   very poorly  poorly  satisfactorily  well  very well

2. On the basis of the information provided, would you recommend this instructor for promotion? (Circle one)

   1  2  3  4  5  
   definitely  probably  uncertain  probably  definitely  
   not         not

3. If this instructor were to explain to you his philosophy about teaching, how likely do you think it is that you would agree with him? (circle one)

   1  2  3  4  5  
   would definitely  probably  neither  probably  definitely  
   disagree     disagree     agree nor disagree

4. Based upon the information provided about this person, how likely do you think it is that you would agree with his beliefs about students? (circle one)

   1  2  3  4  5  
   would definitely  probably  neither  probably  definitely  
   disagree     disagree     agree nor disagree

5. If this instructor's performance changes in the future, in which direction do you think the change will occur? (circle one)

   1  2  3  4  5  
   his performance  his performance  I do not think it  performance  his performance  
   will decline    will decline    will change    will improve
   significantly   a little      significantly   little
APPENDIX 4

DEBRIEFING QUESTIONNAIRE
Name: ____________________________ 

Sex: Male Female (circle one) 

Classification: Freshman ____ Sophomore ____ Junior ____ Senior ____ (check one) 

1. As you were actually rating the instructors, what thoughts went through your mind? 

2. How many of the instructors appeared to you to be actually working? ____________ 

3. To the best of your recollection, how many of the instructors were smiling? ____________ 

4. How many of the instructors wore glasses? ____________ 

5. About how long (on the average) did it take you to fill out a rating form on a given instructor? (check one) 
   less than a minute ______
   about two minutes ______
   about three minutes ______
   about five minutes ______
   about eight minutes ______ 

6. How many of the instructors were sitting at a desk? ______ 

7. What do you think is the most important thing for a college instructor to accomplish when teaching a course?
8. How many of the instructors would you guess could be classified as members of a minority group? ______________

9. How many of the instructors would you say were neatly dressed? ______________

10. If you could have had additional information about each of the instructors, what information would you have wanted most?

11. Did you find some of the instructors easier to rate than others, if so why?

   yes because

   no because
REFERENCES


Distribution List

Navy

4 Dr. Marshall J. Farr, Director Personnel & Training Research Programs Office of Naval Research (Code 456) Arlington, VA 22217

1 ONR Branch Office 495 Summer Street Boston, MA 02210 Attn: Dr. James Lester

1 ONR Branch Office 1030 East Green Street Pasadena, CA 91101 Attn: Dr. Eugene Glove

1 ONR Branch Office 536 S. Clark Street Chicago, IL 60605 Attn: Dr. Charles E. Davis

1 Dr. M. A. Bertin, Scientific Director Office of Naval Research Scientific Liaison Group/Tokyo American Embassy APO San Francisco 96503

1 Office of Naval Research Code 200 Arlington, VA 22217

6 Commanding Officer Naval Research Laboratory Code 2627 Washington, DC 20390

1 Human Resources Program Manager Naval Material Command (0344) Room 1044, Crystal Plaza #5 2221 Jefferson Davis Highway Arlington, VA 20360

1 Special Assistant for Enlisted Force Analysis Bureau of Naval Personnel (Pers 2x) Room 2628, Arlington Annex Washington, DC 20370

1 Director, Human Resource Management Naval Amphibious School Naval Amphibious Base, Little Creek Norfolk, VA 23521

1 LCDR Charles J. Theisen, Jr., MSC, USN 4024 Naval Air Development Center Warminster, PA 18974

1 Commanding Officer U.S. Naval Amphibious School Coronado, CA 92155

1 CDR Paul D. Nelson, MSC, USN Naval Medical R&D Command (Code W4) National Naval Medical Center Bethesda, MD 20014

1 Commanding Officer Naval Health Research Center San Diego, CA 92152 Attn: Library

1 Chairman, Leadership & Law Dept. Div. of Professional Development U.S. Naval Academy Annapolis, MD 21402

1 Scientific Advisor to the Chief of Naval Personnel (Pers Or) Naval Bureau of Personnel Room 4410, Arlington Annex Washington, DC 20370

1 Dr. Jack R. Borsting Provost & Academic Dean U.S. Naval Postgraduate School Monterey, CA 93940

1 Mr. Maurice Callahan ONRAC (Code 2) Dept. of the Navy Bldg. 2, Washington Navy Yard (Anacostia) Washington, DC 20374

1 Office of Civilian Personnel Code 342/02 WAP Washington, DC 20390 Attn: Dr. Richard J. Niehaus

1 Office of Civilian Personnel Code 263 Washington, DC 20390

1 Superintendent (Code 1424) Naval Postgraduate School Monterey, CA 93940

1 Dr. H. M. West III Deputy ADCNO for Civilian Planning and Programming (Acting) Room 2625, Arlington Annex Washington, DC 20370

1 Mr. George N. Graine Naval Sea Systems Command SEA 047C12 Washington, DC 20362

1 Chief of Naval Technical Training Naval Air Station Memphis (75) Millington, TN 38054 Attn: Dr. Norman J. Kerr

1 Principal Civilian Advisor for Education and Training Naval Training Command, Code 00A Pensacola, FL 32508 Attn: Dr. William L. Maloy

1 Dr. Alfred F. Smode, Director Training Analysis & Evaluation Group Department of the Navy Orlando, FL 32813

1 Chief of Naval Education and Training Support (01A) Pensacola, FL 32509

1 Naval Undersea Center Code 303 San Diego, CA 92132 Attn: W. Gary Thomson

1 Navy Personnel R&D Center Code 01 San Diego, CA 92152

5 A. A. Sjoholm, Head, Technical Support Navy Personnel R&D Center Code 201 San Diego, CA 92152

2 Navy Personnel R&D Center Code 310 San Diego, CA 92152 Attn: Dr. Martin F. Wiskoff

1 Dr. Robert Morrison Navy Personnel R&D Center Code 301 San Diego, CA 92152

1 Navy Personnel R&D Center San Diego, CA 92152 Attn: Library

Army

1 Technical Director U.S. Army Research Institute for the Behavioral & Social Sciences 1300 Wilson Boulevard Arlington, VA 22209

1 Armed Forces Staff College Norfolk, VA 23511 Attn: Library

1 Commandant U.S. Army Infantry School Fort Benning, GA 31905 Attn: ATSH-IW-IT

1 Commandant U.S. Army Institute of Administration Attn: EA Fort Benjamin Harrison, IN 46216

1 Dr. Ralph Dusek U.S. Army Research Institute 1300 Wilson Boulevard Arlington, VA 22209

1 Dr. Joseph Ward U.S. Army Research Institute 1300 Wilson Boulevard Arlington, VA 22209

1 Dr. Ralph Canter U.S. Army Research Institute 1300 Wilson Boulevard Arlington, VA 22209

1 Dr. Milton S. Katz, Chief Individual Training & Performance Evaluation Technical Area U.S. Army Research Institute 1300 Wilson Boulevard Arlington, VA 22209

1 HQ USAREUE & 7th Army ODCSOPS USAREUR Director of GED APO New York 09403