Derek L. Mracek, LOCUS OF RATER-RATEE RACE EFFECTS AS INFLUENCED BY RATING SOURCE (Under the direction of Dr. Mark C. Bowler) Department of Psychology, February 2011.

This study examines the impact of both rater and ratee race on job performance ratings. Traditionally, the true nature of race-based distortions to performance ratings is difficult to ascertain due to a lack of true score in performance. By utilizing a series of walk-through performance measures, Cronbach's (1955) accuracy components were used to determine the true nature of race-based distortions. Overall, the majority-member supervisors did not deflate the ratings of minority-members. In fact, the ratings of minority-members were inflated by both source levels; and peers inflated ratings significantly more when compared to supervisors. Moreover, majority members were rated accurately by both supervisors and peers. The implications of this and the potential reasons for it are discussed.

LOCUS OF RATER-RATEE RACE EFFECTS AS INFLUENCED BY RATING SOURCE

A Thesis

Presented to

the Faculty of the Department of Psychology

East Carolina University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts in Psychology

by

Derek L. Mracek

February, 2011

LOCUS OF RATER-RATEE RACE EFFECTS AS INFLUENCED BY RATING SOURCE

by

Derek L. Mracek

APPROVED BY:

DIRECTOR OF THESIS	Mark C. Bowler, Ph.D.
COMMITTEE MEMBER	William F. Grossnickle, Ph.D.
COMMITTEE MEMBER	Karl L. Wuensch, Ph.D.
CHAIR, DEPARTMENT OF PSYCHOLOGY	Kathleen A. Row, Ph.D.
DEAN OF THE GRADUATE SCHOOL	Paul J. Gemperline, Ph.D.

Acknowledgements

I would like to sincerely thank the director of thesis, Dr. Mark C. Bowler, for his guidance, patience, and resolve; as well as Dr. William F. Grossnickle and Dr. Karl L. Wuensch for their patience, support, and attention to detail.

Table of Contents

List of Tables

CHAPTER I: INTRODUCTION

The process of performance evaluation – generating data describing employee behavior – includes identifying, observing, measuring, and evaluating employee behaviors (Carroll & Schneir, 1982). This information is subsequently used to plan, organize, direct, and control employee talent within organizations (Fayol, 1949). At the foundation of this system are the assumptions that (1) accurate levels of performance can be assessed (Ferris & Judge, 1991; Ones, Viswesvaran, & Schmidt, 2008; Woehr, 2008) and that (2) these assessments can be used to differentiate individual performance (Cardy & Dobbins, 1994). However, due to the complex and intertwined nature of organizational structure, only a limited number of situations allow for objective indices to be utilized as the primary source of data for performance evaluations (Cardy & Krzystofiak, 1991; Murphy & Cleveland, 1995). Moreover, most jobs are comprised of tasks that evolve over time, rendering objective indices even more difficult to attain (Osterman, 2007). Therefore, subjective ratings by supervisors and peers are commonly utilized to evaluate all aspects of job performance (Schmidt, 1988). Unfortunately, subjective ratings are plagued by a host of problems including, but not limited to, halo, leniency, severity, central tendency, purposeful manipulation, and bias from race, gender, or age (Cardy & Dobbins, 1994; Landy & Farr, 1980; Saal, Downey & Lahey, 1980).

The primary issue that arises from this host of problems inherent in performance evaluations is the accuracy of the judgments that are made (Murphy, 2008). Specifically, the relationship between the evaluations of those making the subjective ratings, the raters, and the true performance of those being rated – the ratees – is

uncertain (Murphy, 2008; Woehr, 2008). These judgments may be extremely inaccurate and biased by the demographics of both the rater and the ratee (Carroll et al., 1982). In particular, majority- and minority-race raters appear to use different cues when evaluating the performance of majority- and minority-ratees (Pulakos, Oppler, White, & Borman, 1989; Stauffer & Buckley, 2005). In turn, this leads to differential subjective judgments being made regarding the performance of employees (Stauffer & Buckley, 2005).

The impact of racial bias on performance evaluation has been argued to be one of the prominent reasons why disparate treatment still continues in personnel decisions (Maass, Castelli, & Arcuri, 2000; Osterman, 2007). The primary explanation for this is that most raters enter the evaluation process with well-developed performance-related schema for members of different racial groups (DeNisi, Cafferty, & Meglino, 1984; Feldman, 1986). Furthermore, the difficulties commonly associated with making performance ratings (e.g., information overload; opportunity to observe) influence whether or not stereotypes are manifested and the degree to which the stereotype affects the accuracy of the evaluation (Funder, 1987; Johnson & Cochran, 2008).

To determine whether different standards are being applied to different groups, the accuracy of ratings must be assessed. This allows for the identification of any systematic difference between the ratings of members of different demographic groups (Pulakos et al., 1989; Rotundo & Sackett, 1999). Accuracy of ratings can only be determined when the true level of performance is known; however, true performance levels are generally unavailable (Woehr, 2008). This inability to determine the accuracy of performance evaluations has been predominately due to the inherent difficulty

separating true ratee performance from rater error (Pulakos et al., 1989). Without knowing the true level of ratee performance it is not clear whether differences in how an individual is rated by members of different demographic groups is due to leniency by one rater, severity by the other, or a combination of the two (Rotundo & Sackett, 1999; Sackett & DuBois, 1991). In the present study, a normative standard of behavior was utilized to help clarify this issue. Subsequently, this locus and direction of racial bias can be ascertained and a better understanding of rater-ratee race effects can be determined. This racial bias, or *rater-ratee race effect*, is affected by the rating source of the evaluation. The present study sought to better understand the rater-ratee race effect and to investigate the conditions where the rater-ratee race composition influences the accuracy of an evaluation.

Understating the Impact of Race on Performance Ratings

There is much theory that posits that ratings can be adversely influenced by a raters' stereotypes and are not a reflection of the ratee's actual performance behavior (Greenwald, 2008). More specifically, schemas are well-developed knowledge structures that provide cognitive short cuts for organizing characteristics, behaviors, or other information on some defining characteristic (e.g., race, sex, or age) or construct (e.g., performance). Stereotypes are a specific type of schema regarding people on a defining characteristic. Regardless of the specificity, these knowledge structures possess a cluster of associations to organize information (Dobbins, Cardy & Truxillo, 1988). The performance schema, or implicit theory of performance, is characterized by a cluster of associations regarding job tasks and effective or ineffective behaviors that include, but are not limited to, objective indices of job performance (e.g., productivity,

absenteeism, turnover), job-related knowledge (e.g., education, training), job-related skills (e.g., writing, time management, instructing), and job-related abilities (e.g., reasoning, general mental ability, oral expression; Hastie, 1981; Peterson, Borman, Hanson, Kubisiak, 1999).

The categorization of a ratee as a high or low performer and the subsequent social cognition mechanisms (e.g., categorization, observation, recall) are biased by negative stereotypes of minority individuals. Raters have a tendency to associate the minority-group ratee individuals with their low performance schema and are less likely to associate these individuals with their high performance schema. (Dobbins et al., 1988; Hastie, 1981). Raters often have not observed enough behavior on which to categorize and subsequently base an evaluation. In theory, raters are relegated to their negative stereotypes to extrapolate beyond the observed behavior to infer ratees' performance (Bielby, 2000; Neisser, 1967). That is, a minority ratee is categorized as a low performer by default and this categorization unfavorably guides expectations of the rater and influences the rater's cognitive processes (Alba & Hasher, 1983). Furthermore, these fallible social cognition processes are exacerbated by typical constraints characteristic of the rating situation (e.g., information overload, deadlines) – all of which increase the difficulty to accurately recall ratee behavior (Barnes-Farrell, 2001). A negative stereotype would be manifested in a performance evaluation; wherein, raters rate a minority individual lower than their true performance. This process of a negative stereotype being manifested in an adverse performance evaluation has not been empirically supported in an applied setting (Landy, 2008).

Rater-Ratee Race Effects. Majority-group supervisors understating the performance of minority-group ratees is a salient possibility for why mean racial differences in performance evaluations exist (Nkomo & Cox, 1989; Stauffer & Buckley, 2005). Supervisor raters with active negative stereotypes could deflate their ratings; however, it has also been argued that inaccurate performance ratings can favor minority members. Under affirmative action and equal employment pressures, race is likely to be an important consideration potentially motivating supervisors to inflate their ratings of minority-group members to reduce the likelihood of legal scrutiny associated with low ratings (Kraiger & Ford, 1985). This could lead a supervisor to inflate his or her supervisor rating favoring minority ratees (Kraiger & Ford, 1985; Tetlock et al., 2008).

Many studies show significant race effects, but cannot separate the relative contributions of ratee performance and rater effects to rating differences (Kraiger & Ford, 1985; Landy & Farr, 1980; Sackett & Dubois, 1991). As is the case, a clear explication of race effects and the possible interpretation as racial bias is complex (Kraiger & Ford, 1985; Pulakos et al. 1989; Sackett & DuBois, 1991; Staufer & Buckley, 2005). Kraiger and Ford (1985), paralleling the previous work of Landy and Farr (1980), supported the notion that ratees tend to receive higher supervisor ratings from raters of the same race. Majority-group raters assigned significantly higher ratings to majority-member ratees than to minority-group ratees and minority-group raters assigned significantly higher ratings to minority-group ratees than to majority-group ratees. Furthermore, this effect was equally strong for both majority- and minority-group members. Kraiger and Ford concluded that raters evaluated many of the same ratees, a logical conclusion is that ratings were biased to some degree. However, because the

meta-analysis utilized a between-subjects ANOVA design, wherein, a ratee was not evaluated by both a majority- and minority-group rater, actual performance differences between races could have existed in the sample. Thus, racial bias cannot be ascertained; however, Kraiger and Ford suggested that it is most likely due to a combination of bias and performance differences.

In a separate meta-analytic review, Sackett and Dubois (1991) challenged the conclusions of Kraiger and Ford (1985) by utilizing the analysis of rater race as a within-subjects factor. A within-subjects design eliminates the potential issue of true differences in performance by having both a majority- and minority-group rater evaluate the same individual. Thus, any differences in ratings can be unambiguously attributed to rater differences. The analysis supported that majority-group ratees received higher ratings than minority-ratees from both minority- and majority-supervisor raters, although the magnitude of the difference was substantially larger for majority-member raters. Sackett and Dubois (1991) concluded that minority-group raters rate majority-group ratees slightly higher than they rate minority-group ratees. Thus, minority-group members do not rate members of their own race higher as previously asserted by Kraiger & Ford and Landy & Farr (1980). Furthermore, majority-group ratees received similar ratings from both minority- and majority-group raters, whereas minority-group ratees received higher ratings from minority-group raters than from majority-group raters. This interaction effect; namely, rater-ratee race effect is the primary concern – majority- and minority-raters differed in their ratings of minority-ratees. As Sackett and DuBois (1991) stated:

> "Black ratees received ratings from white raters that ranged from .02 to .10 of a standard deviation lower than the ratings that they received from black raters. In contrast, ratings for white ratees from Black raters ranged from .003 of a standard deviation lower to .003 of a standard deviation higher than ratings from white raters" (p. 876).

Sackett and Dubois noted that minority- and majority-supervisors disagree on the magnitude of the minority-ratees' performance; however, it is unknown whether this is a function of minority-raters inflating their ratings, majority-supervisors deflating the performance of minorities, or a combination of both. Furthermore, Sackett et al. (1991, 1999) have noted that there is no accepted method of establishing whether there is bias in performance ratings, but assert that the difference between rated performance and true performance must be systematically larger for members of one group than for another, and that asserting it is bias is not tenable because true performance is not known.

Stauffer and Buckley (2005) disputed the conclusions of Sackett and Dubois (1991) and noted that even a trivially significant rate-race interaction effect constitutes racial bias and is practically significant. More specifically, the supervisor race effect reflects a source of bias and is defined by the difference between the column marginal means (see Table 1). The difference in direction of mean group performance between races would support a true difference (McKay & McDaniel, 2006; Schmidt, 1988) if the within-group effect (i.e., supervisor race) was zero. Moreover, the rater-ratee race interaction effect suggests that the two groups of supervisors do not agree on the magnitude. Thus, a significant interaction indicates a systematic distortion of true

performance relationships by one or both rater groups. In this case, if you are a minority-group ratee it is important whether your supervisor is a minority- or majority-group rater.

Table 1. *Typical Rater X Ratee Race Effects Study Design*

Ratee	**Majority Rater**	**Minority Rater**	
Minority	M_{WW}	M_{WB}	$M_{W.}$
Majority	M_{BW}	M_{BB}	$M_{B.}$
	$M_{.W}$	$M_{.B}$	

Note. The first letter indicates the mean; the first letter of the subscript denotes rate race (W = Majority member, B = Minority member); the second letter of the subscript denotes rater race. Marginal mean values are indicated with a dot.

The typical rater-ratee interaction effect is small (e.g., d = .05); however, the practice of translating race-related differences into metrics that are easily understood ignore the practical significance of small effect sizes producing large effects (Eagly, 1995; Greenwald, 2008; Martell, Lane, & Emrich, 1996; Stauffer & Buckley, 2005). More specifically, the importance of a difference depends on the consequences of the rater-ratee race effect (Eagly, 1995). There is considerable support that a small interaction effect in performance ratings would lead to substantially lower promotion ratings for minority-group ratees (Martell et al., 1996). Consider a hierarchical organizational structure and the likelihood that early career success is a critical factor for subsequent promotion, initial performance ratings strongly influence the likelihood of whether or not an individual advances in the organization (Rosenbaum, 1979). Consequently, there would be proportionately fewer minority-group members than majority-group members

at the top levels of an organization; and a subsequent, lack of senior level mentors of the same race for minority-group members (Ilgen & Youtz, 1984).

Rating Accuracy. It is impossible to discern from the previous data whether the rater-ratee race effect results from majority-member supervisors understating the performance of minority-members, minority-supervisors overstating it, or a combination of both of these factors (Stauffer & Buckley, 2005). In the present study, to better understand whether raters' evaluations are correct, under-correct, or over-correct the accuracy of ratings were ascertained by making comparisons of performance ratings against a standard that indicates the actual performance levels. More specifically, with a normative score of performance, Cronbach's accuracy components (1955) can be calculated (Cardy & Dobbins, 1994).

Cronbach (1955) decomposed a squared difference measure of accuracy into four components, which is widely accepted as a comprehensive means of assessing ratings accuracy (Dobbins et al. 1988; Roach & Gupta, 1992) When assessing the relationships of these components, Roach and Gupta (1992) noted that each component contributes unique information about rating accuracy; hence, rating accuracy is a multidimensional construct. These components are: elevation, differential elevation, stereotype accuracy, and differential accuracy. For the current student, elevation, and differential elevation was utilized to test our hypotheses. The components of stereotype accuracy and differential accuracy were not computed, because these components require information on scores for more than two dimensions of performance. The present study made comparisons of performance ratings against a standard that indicates the normative performance levels for each of the following two

dimensions: technical knowledge and technical proficiency. Elevation is a measure of the tendency to evaluate a group of ratees as too high or too low with respect to the average true level of performance exhibited by the ratees. Differential elevation indicates the accuracy of differentiation among ratees, controlling for the overall rating tendency of the rater; essentially, this is the accuracy that a rater differentiates among average (across dimensions) ratee performance levels. The two accuracy components were calculated on the basis of the formulae presented by Murphy, Garcia, Kerkar, Martin, and Balzer (1982) and Cardy and Dobbins (1994). For a rater who evaluates *n* ratees on *k* items or dimensions, scores on elevation (EL) and differential elevation (DE) are given by the following terms:

$$E^2 = (\bar{r}.. - \bar{t}..)^2$$

$$DE^2 = \frac{1}{n}\sum_i [(\bar{r}_{i.} - \bar{r}_{..}) - (\bar{t}_{i.} - \bar{t}_{..})]^2$$

Where

$r_i.$ and $t_i.$ = mean rating and true score for ratee i,

$r..$ and $t..$ = mean rating and true score over all ratees and items.

An additional measure of rating accuracy was included in the study (Cardy & Dobbins, 1994).The present study calculated simple difference scores that are directional analogs of elevation. These variance scores result in small scores for the elevation analog (Murphy & Balzer, 1989); wherein, a simple difference measure would provide a direct test of a directional hypothesis to determine significant within-group effects on a group of ratees (Cardy & Dobbins, 1994). That is, Cronbach's accuracy components are squared deviation measures; they reflect the amount of error magnitude, but do not ascertain the direction of the rater error (e.g., inflation or deflation of ratings); and

subsequently the locus of the rater-ratee race effect. In the present study, the direction of the rater error is integral to testing the hypotheses, thus, the variance score versions of the formulae and not the correlational versions were utilized to determine the absolute amount of error variance in ratings (Becker & Cardy, 1986). In turn, the unsquared elevation analog allows for an analysis of both the magnitude and direction of rater error (e.g., inflation or deflation of ratings). For example, if majority-group members are rated three units higher and minority-group members are rated three units lower than their normative score, both sets of ratings would be equally inaccurate from a magnitude perspective, and analyses of the accuracy component would not detect any difference. Thus, in order to detect directional shifts in performance ratings, the average (unsquared) deviation between ratings and normative levels of performance will be included as an additional accuracy measure. It is hypothesized that minority- and majority-member supervisors will not rate minority-subordinates similarly, resulting in a rater-ratee race effect for subordinate ratings (Sackett et al. 1991; Stauffer et al., 2005). Specifically:

H1: Majority-member supervisors will provide accurate ratings (EL) of majority-member subordinates.

H2: Majority-member supervisors will provide accurate ratings (DE) of majority-member subordinates.

H3: Minority-member supervisors will provide accurate ratings (EL) of majority-member subordinates.

H4: Minority-member supervisors will provide accurate ratings (DE) of majority-member subordinates.

H5: Majority-member supervisors will deflate (EL) ratings of minority-subordinates

H6: Majority-member supervisors will differentially elevate (DE) ratings of minority-subordinates.

H7: Minority-supervisors will provide accurate ratings (EL) of minority-subordinates.

H8: Minority-supervisors will provide accurate ratings (DE) of minority-subordinates.

H9: Majority-supervisors will rate minority-subordinates lower (EL) than majority-member subordinates.

H10: Minority-supervisors will rate minority-subordinates higher (EL) than majority-member subordinates.

Source of Rating. An additional concern with rater-ratee race effects relates to the source of the ratings being made. In most organizations, decisions are not made by group members but by individuals who are usually in a superior position to the individual being rated (Landy, 2008). However, with the increase of team-based structures, peer ratings are seeing an increased use in organizations (cf., Borman, 1997; Facteau & Craig, 2001; Harris & Schaubroeck, 1988; Norman & Zawacki, 1991). Furthermore, peer evaluations have been shown to be salient predictors of potential job performance (e.g., Kane & Lawler, 1978; Reilly & Chao, 1982) and peer ratings have also been shown to predict job advancement more effectively than assessment center ratings (Shore, Shore, and Thornton, 1992). Peer ratings have also been used to develop consensus (Harris et al., 1988), improve legal defensibility of evaluation systems (Bernardin &

Beatty, 1984), and increase reliability, fairness, and ratee acceptance of evaluation systems (Latham & Wexel, 1982).

The utility of using multiple sources of raters is that one rater may not have sufficient opportunity to observe job performance in order to attain a well-rounded evaluation of ratees' performance dimensions (Ivancevich, 2001). Utilizing multiple sources provides an enhanced ability to sufficiently observe various job facets leading to fairer and more accurate judgments (Borman, 1974; Henderson, 1984; Latham & Wexel, 1982). Moreover, peers may be in a better position and are more knowledgeable of the behaviors that are critical for favorable job performance and observe more work behavior; whereas, supervisors rely on work outcomes (Carson, Cardy, & Dobbins, 1991). Observing more work behavior, peers are more likely to differentiate effort from performance, and are more likely to focus on task-relevant factors (Klimoski & London, 1974; Murphy & Cleveland, 1995). Furthermore, peer ratings may be less susceptible to stereotypes as they do not have to exercise as much mental resources to recreate actual behaviors that have been exhibited by the ratee (Cardy & Dobbins, 1994). Furthermore, peers are more likely to form a personal relationship with their coworkers preventing against the effects of positive or negative stereotypes; subsequently, ratings of minority groups should be more accurate than supervisor ratings (Landy, 2008).

H11: Peer raters will provide more accurate ratings (EL) when compared to supervisor raters.

.

CHAPTER II: METHOD

High Fidelity Setting

Germane to the relationship of raters differentially evaluating minority- and majority-group members is the setting of the research sample. One of the major dichotomies in research in performance evaluations is the defining characteristics of the laboratory and field settings (Landy, 2008; Murphy, 2008). Research employing field samples examines relationships within a close proximity to the situational context to which they are to be generalized (Funder, 1987). In contrast, research employing laboratory samples examine basic psychological processes and are characterized as artificial and unnatural. Whether or not stereotypes are activated and influence personnel decision making depends on a combination of cognitive and motivational variables and mechanisms (Kunda & Spencer, 2003; Maynard & Brooks, 2008). Field samples include a dynamic factor of behavior where a trained accountable individual is evaluating a subordinate or peer under actual job conditions where administrative consequences affect the rater or ratee, and this dynamic is not found in research samples (Landy, 2008). Moreover, laboratory research regarding race is overwhelmed by a participants desire to give socially desirable responses (Kraiger & Ford, 1985). Therefore, replicating behavior relevant to racial bias in laboratory research is highly questionable, a primary reason why it has been argued that investigating racial bias is more relevant and meaningful to a setting of administrative personnel decisions (Copus 2005; Landy, 2008; Tetlock et al., 2008). Interestingly, these two unique methodologies have also engendered somewhat different results: racial bias is more likely to be found

in field than laboratory settings (Ford, Kraiger, Schechtman, 1986; Kraiger & Ford, 1985).

A primary reason why the two unique methodologies engender different results is that distinct memory systems are utilized for each setting. Implicit and explicit memory processes affect evaluations differently (Lord & Maher, 1991). This is a primary reason why differences in the relationship of race effects in laboratory studies and field studies can be attributed to the distinct memory systems utilized: laboratory ratings are based on explicit, or consciously retrieved knowledge; whereas, implicit knowledge is represented in the schemas or stereotypes (Lord & Maher, 1991). In theory, laboratory research involves more controlled processing than in typical field situations because raters are evaluating individuals they are not familiar with and may be able to focus all of their attention on the ratee's performance (Cardy & Dobbins, 1994). In a field setting, a supervisor forms a personal relationship with a subordinate and the mean level of perceived ability to perform of that individual's demographic group could be altered (see Landy, 2008). However, raters may not have ample opportunity to form a personal relationship and experience the subsequent individuating information. Furthermore, stereotyping not only affects the end result: evaluations, but also affects the performance of the stereotyped individual (Wessel & Ryan, 2008), as these individuals are inherently stigmatized individuals in selection (Word, Zanna, & Cooper, 1974). Overall, these effects can be framed within controlled or implicit memory systems. This may provide a bridge to understanding rater-race effects as the type of memory system may be more predominant in certain settings where evaluation data is collected. Namely, implicit memory systems are more characteristic of field settings, factoring in

the nature of performance evaluation: lack of observing ratee behavior, information overload, imperfect recall, and deadlines (Barnes-Farrell, 2001). Furthermore, when predicting the potential performance of candidates raters are making inferences about future, unknown performance, which are more likely to rely on stereotypes (Dobbins, Cardy & Truxillo, 1988). Thus, inaccurate evaluations are more likely to be manifested in performance ratings used for administrative purposes (Dobbins et al., 1988).

The complex nature of racial bias may make an artificial setting more useful in advancing understanding of basic level social psychology processes; however, it does not provide a basis for making inferences about the nature of stereotyping or discrimination and its subsequent effect on performance evaluation (Copus, 2005). This is why research on performance ratings should come from actual evaluations made for a legitimate administrative purpose, such as organizational performance ratings or promotions (Johnson & Cochran, 2008). A field sample possesses representativeness to the population and can identify boundary conditions racial bias may be occurring or may be more susceptible to occur (Cardy & Dobbins, 1994; Landy, 2008). Thus, the current study utilized actual evaluations made for an administrative purpose to describe the rater-ratee race effects in the most meaningful context of administrative personnel decisions (Landy, 2008).

Participants

Of the 1,035 incumbents in the joint-services (Wigdor & Green, 1991) and USAF JPM project (see Hedge & Teachout, 1986, 1992; Ree, Earles, & Teachout, 1994) 945 participants were utilized for the current study. The data were collected for eight jobs as part of the joint-services (Wigdor & Green, 1991) and USAF JPM project (Hedge &

Teachout, 1986, 1992). These personnel represented eight Air Force specialties: Aerospace Ground Equipment Mechanic, Aircrew Life Support Specialist, Air Traffic Control Operator, Avionic Communications Specialist, Information Systems Radio Operator, Jet Engine Mechanic, Personnel Specialist, and Precision Measurement Equipment Laboratory Specialist.

Participants entered service from 1984 through 1988 and completed both basic military training and technical training during their first term of enlistment. Incumbents are 17- to 23-year-old graduates of high school or better (99.1 percent), with average job tenure of about 28 months. Majority-members (Caucasians) consist of 77.9 percent of the ratees and minority-members (non-Caucasians) compromise the remaining 22.1 percent. Men consist of 82 percent of the ratees; while women consist of 18 percent. Majority-members consist of 76.4 percent of the supervisor raters and minority-members compromise the remaining 23.6 percent. Men consist of 90.1 percent of the supervisor raters; while women consist of 9.9 percent. Majority-members consist of 78.8 percent of the peer raters and minority-members compromise the remaining 21.2 percent. Men consist of 83.2 percent of the peer raters; while women consist of 16.8 percent.

Each participant was rated once by each supervisor and 1-3 times by their peers'. For the 945 supervisor ratings, 52 supervisors rated more than one ratee. There was a notable amount of variability with respect to how many times each ratee was rated by their peers: (a) 65 ratees were rated by one peer; (b) 340 ratees were rated by two peers, and (c) 540 ratees were rated by three peers. Similarly, there was a

significant number of peer raters that rated more than one individual. For the 2323 peer ratings, 612 peer raters (26.3 percent) rated more than one ratee.

Measures

The job performance measurement system included hands-on performance tests (HOPTs), interview work sample tests (INT), and a walk-through performance test (WTPT; Hedge & Teachout, 1992). These functioned as the true score of participant performance when computing Cronbach's accuracy components. The HOPTs and INTs were constructed for each job to assess proficiency on representative job tasks and required the examinee to accomplish the tasks, either manually or verbally, at the work site under the observation of a trained administrator, who scored each step with a dichotomous yes-no format. Performance on some tasks was measured only with hands-on work samples, where incumbents were instructed to perform each task according to standardized descriptions of work procedures required for successful task completion. Every task in the WTPT, roughly 20 to 30 for each job, had a maximum time limit. The time limits were established by subject matter experts so that examinees could perform each task without time pressure. Performance was measured on the other tasks with interview testing for situations in which hands-on testing was not feasible because of practical constraints. Thus, when appropriate, interview testing was used to measure all aspects of the performance domain. Performance on other tasks was measured only with interview work samples, where incumbents were required to describe the steps necessary for task completion in a demonstrational manner without the aid of technical manual information. For each job, tasks unique to both the hands-on or interview format, and several tasks overlapped between formats. Each WTPT task

was composed of a series of steps that had been previously weighted on the basis of the importance of that step to the successful completion of the task. These task scores were computed as the percentage of steps completed correctly. The task domains for each job are identified and defined from the Air Force Occupational Survey database (Christal, 1974), which is administered approximately every four years to keep the job content domains current.

For each task, work sample developers used U.S. Air Force technical orders and manuals as well as input from subject matter experts to define and describe the procedure required for successful task completion. The work-sample tests were constructed for each task, reviewed by subject matter experts, and tested at several air force bases across the world. The work-sample tests were administered to the examinees and scored by officers with extensive job experience. The administrators received 1-2 weeks of observation and scorer accuracy training aimed toward increasing inter-rater reliability and accurate observation of examinee completion of task steps (Hedge, Lipscomb, & Teachout, 1988). Administrators received intensive instruction on recognizing instances of correct and incorrect performance on work-sample task steps in videotaped performances. These procedures have been shown to produce accurate and reliable work sample test ratings (Hedge, Dickinson, & Bierstedt, 1988) as reflected in the high average agreement and high average correlational accuracy between their ratings and videotape target ratings.

Job Performance Ratings

In addition to hands-on performance tests and interview work-sample tests, job performance ratings, experience, and demographic information were also gathered from

job incumbents, supervisors, and peers (Hedge & Teachout, 1992). Four rating forms were utilized that measure job performance for varying degrees of specificity: task, dimensional, Air Force-wide, and global. Ratings were made with a paper-and-pencil 5-point, adjectivally anchored scale, by job incumbents, supervisors, and up to three peers, for all eight specialties. Performance was rated for each task on a 5-point scale ranging from 1 *(never meets acceptable level of proficiency)* to 5 *(always exceeds acceptable level of proficiency).* In addition to the task- and dimension-level, performance ratings will be utilized for aspects of proficiency that were assumed to be generalizable across different Air Force Specialties.

Computing the Accuracy Components

Ratings of technical proficiency, technical knowledge, and the walk-through performance test were used to compute Cronbach's Evaluative Accuracy Components (Cronbach, 1955). To create the differential elevation accuracy component the mean rating and mean true score over all ratees and dimensions (e.g., technical knowledge and technical proficiency) for each rater was computed and converted into standard scores. Also, each ratee's mean rating and true score were computed and converted into standard scores. These two parts of the accuracy components were then transformed by adding 3.5 to each. This transformation was performed to avoid any confounding effects due to the order of operations. The differential elevation component was only computed for cases when both – raters rated more than one ratee, and ratees were rated by more than one rater (for each level, supervisor or peer).

CHAPTER III: RESULTS

H1: Majority-member supervisors will provide accurate ratings (EL) of majority-member subordinates

The mean elevation accuracy component analog score (M = -.10, s = 1.29) for majority-member supervisors rating majority-member subordinates was not significantly different than the accuracy rating ideal of zero (0), $t(572) = 1.85$, $p = .07$, $d = .08$. See Table 2 for descriptive statistics of all rater-ratee combinations. The cell means did not deviate significantly from zero: it can be inferred that majority-member supervisors do not elevate (EL) ratings of majority-member subordinates. This evidence supports hypothesis one that majority-member supervisors will provide accurate ratings (EL) of majority-member subordinates. In fact, majority-member supervisors did not deviate significantly from the rating ideal of zero when rating majority-member subordinates (EL).

H2: Majority-member supervisors will provide accurate ratings (DE) of majority-member subordinates

A one-sample *t*-test was employed utilizing the differential elevation accuracy component as the dependent measure. The differential elevation component utilized is a squared term, hence only the magnitude of the effect can be ascertained. The mean differential elevation accuracy component score (M = .18, s = .22) for majority-member supervisors rating majority-member subordinates was significantly different than the accuracy rating ideal of zero (0), $t(31) = 4.7$, $p < .01$, $d = .84$. The cell means did deviate significantly from zero. This evidence fails to support the hypothesis two that majority-member supervisors will perform accurate ratings (DE) of majority-member

subordinates. In fact, majority-member raters did differentially evaluate (DE) majority-member subordinates. That is, majority-member supervisors did deviate significantly from the rating ideal of zero when rating majority-member subordinates (DE).

H3: Minority-member supervisors will provide accurate ratings (EL) of majority-member subordinates

The mean elevation accuracy component analog score (M = -.01, s = 1.29) for minority-member supervisors rating majority-member subordinates was not significantly different than the accuracy rating ideal of zero (0), $t(161) = .11$, $p = .91$, $d = .01$. The cell means did not deviate significantly from zero. Thus, this evidence supports hypothesis three: minority-member supervisors will perform accurate ratings (EL) of majority-member subordinate**s.**.

H4: Minority-member supervisors will provide accurate ratings (DE) of majority-member subordinates

The mean differential elevation accuracy component score (M = .28, s = .19) for minority-member supervisors rating majority-member subordinates was significantly different than the accuracy rating ideal of zero (0), $t(9) = 2.36$, $p < .01$. The cell means did deviate significantly from zero, thus indicating minority-member supervisors did differentially elevate (DE) ratings of majority-member subordinates. This evidence fails to support hypothesis four that minority-member supervisors will perform accurate ratings (DE) of majority-member subordinates. In fact, minority-member supervisors did differentially elevate (DE) ratings of majority-member subordinates.

H5: Majority-member supervisors will deflate (EL) ratings of minority-subordinates

A one-sample *t*-test was employed utilizing the elevation accuracy component analog as the dependent measure. The rater-ratee combination of majority-member raters rating minority-member ratees elicited a significant difference from the rating ideal of zero (0). The mean elevation accuracy component analog score (M = .23, s = 1.46) for majority-member supervisors rating minority-member subordinates was significantly different than the accuracy rating ideal of zero, $t(148) = 1.94$, $p = .05$, $d = .16$. That is, the cell means did deviate significantly from zero, as majority-member supervisors inflated minority-subordinates when compared to the accuracy rating ideal of zero. This evidence does not support hypothesis five that majority-member supervisors will deflate (EL) ratings of minority-subordinates. In fact, the relationship was in the opposite direction that was hypothesized: majority-member supervisors inflated minority-subordinates when compared to the accuracy rating ideal of zero.

H6: Majority-member supervisors will differentially elevate (DE) ratings of minority-subordinates

The mean differential elevation accuracy component score (M = .20, s = .28) for majority-member supervisors rating minority-member subordinates was significantly different than the accuracy rating ideal of zero (0), $t(9) = 2.36$, $p < .04$. The cell means did deviate significantly from zero, thus indicating majority-member supervisors did differentially elevate (DE) ratings of minority-member subordinates. This evidence supports hypothesis six that majority-member supervisors will differentially elevate (DE) ratings of minority-member subordinates. In fact, majority-member supervisors elevated minority-member subordinates (EL) and also majority-member supervisors differentially elevated (DE) ratings of minority-member subordinates.

H7: Minority-supervisors will provide accurate ratings (EL) of minority-subordinates

The mean elevation accuracy component analog score (M = .30, s = 1.17) for minority-member supervisors rating minority-member subordinates was not significantly different than the accuracy rating ideal of zero (0), $t(59) = 1.95$, p = .06, d = .25. The cell means did not deviate significantly from zero. This evidence supports hypothesis seven that minority-supervisors will perform accurate ratings (EL) of minority-subordinates. That is, minority-member raters did not elevate (EL) ratings of minority-member ratees. In conjunction with the support of hypothesis three, minority-member supervisors performed accurate ratings (EL) of both minority- and majority-subordinates.

H8: Minority-supervisors will provide accurate ratings (DE) of minority-subordinates

The mean differential elevation accuracy component score (M = .14, s = .10) for minority-member supervisors rating minority-member subordinates was significantly different than the accuracy rating ideal of zero (0), $t(4) = 3.21$, p = .03, d = 1.61. The cell means did deviate significantly from zero, thus indicating minority-members differentially elevate (DE) ratings of minority-member ratees. This evidence fails to support hypothesis seven that minority-supervisors will perform accurate ratings (DE) of minority-subordinates. In fact, minority-supervisors did differentially elevate (DE) ratings of minority-subordinates.

H9: Majority-supervisors will rate minority-subordinates lower (EL) than majority-member subordinates

The combination of supervisor rater-ratee race significantly affected elevation accuracy component analog scores, $F(3, 940) = 3.64$, $p < .01$, $\eta^2 = .011$, 90% CI [.001, .023]. The Bonferroni procedure was used to conduct pair-wise comparisons holding family-wise error at a maximum of .05 (see Table 3). The elevation component score for majority-supervisors rating minority-subordinates ($M = -.01$, $s = 1.32$) was not significantly greater than the elevation component score for majority-supervisors rating majority-subordinates ($M = -.10$, $s = 1.29$) ($p = 1$). This evidence fails to support hypothesis nine that majority-member supervisors will rate minority-subordinates lower (EL) than majority-member subordinates. In fact, majority-member supervisors did not rate minority-subordinates lower (EL) than majority-member subordinates.

H10: Minority-supervisors will rate minority-subordinates higher (EL) than majority-member subordinates

The combination of supervisor rater-ratee race significantly affected elevation accuracy component analog scores, $F(3, 940) = 3.64$, $p < .01$, $\eta^2 = .011$, 90% CI [.001, .023]. The Bonferroni procedure was used to conduct pair-wise comparisons holding family-wise error at a maximum of .05 (see Table 3). The elevation component score for minority-supervisors rating minority-subordinates ($M = .30$, $s = 1.17$) was not significantly greater than the elevation component score for majority-supervisors rating minority-subordinates ($M = -.01$, $s = 1.32$) ($p = .74$). Subsequently, this evidence fails to support hypothesis ten that minority-supervisors will rate minority-subordinates higher (EL) than majority-member subordinates. In fact, minority-member subordinates did not rate minority-group member subordinates significantly higher than majority-member subordinates.

H11: Peer raters will provide more accurate ratings (EL) when compared to supervisor raters

Comparing the magnitude and direction of error across rater levels (e.g., supervisor, peer) tests hypothesis eleven that peer raters will less likely elevate (EL) ratings of peer ratees compared with supervisor raters. A one-way MANOVA was utilized with supervisor and peer elevation accuracy components scores functioning as the within-subjects factor to test the aforementioned hypothesis. Ratee race functioned as the between-subjects factor. A significant effect was found for the main effect of Rater Level, $F(1, 2048) = 4.02$, $p = .05$, Roy's Largest Root = .002, 90% CI [.00, .06]. Overall, peer raters were more likely to elevate ratings than supervisor raters. Thus, peers are more lenient than supervisors when rating the same ratees. A significant effect was found for the interaction effect of Rater Level X Ratee Race, $F(1, 2048) = 5.27$, $p = .02$, Roy's Largest Root = .003, 90% CI [.000, .008]. Pillai's Trace and Roy's Largest Root were the same value. Importantly, while supervisor raters elevate minority-subordinates', minority and majority-member peer raters significantly inflate minority-ratees more than supervisor raters do. Moreover, when assessing the accuracy of majority-ratees, there is no difference across rater race at both the supervisor and peer rater level with respect to direction and magnitude of elevation. The current study rejects the hypothesis that peer raters will perform more accurate ratings (EL) when compared to supervisor raters. Regardless of Rater Race, both peer and supervisor raters inflate the performance of minority-member ratees; peers more so when it comes to minority-ratees.

Table 2. *Group Mean Deviation Elevation Accuracy Component Analog*

Rater-Ratee Race	**Difference**		***SEM***		***df***		***t***		***p***		***CI*.95**	
	Sup	Peer	Sup	Peer	Sup	Peer	Sup	Peer	Sup	Peer	Sup	Peer
Majority Rater Majority Ratee	-.10	-.09	.05	.04	572	1262	1.85	2.70	.07	.01**	(-.20) – (.01)	(-.16) – (-.02)
Minority Rater Minority Ratee	.30	.44	.15	.11	59	105	1.95	3.80	.06	.00**	(-.08) – (.60)	(.21) – (.66)
Minority Rater Majority Ratee	-.01	-.13	.10	.07	161	328	.11	-1.84	.91	.07	(-.22) – (.19)	(-.27) – (.01)
Majority Rater Minority Ratee	.23	.35	.12	.07	148	354	1.94	5.26	.05*	.00**	(-.00) – (.47)	(.22) – (.48)
Total	-.01	.00	.00	.03	943	2052	.03	.15	.07	.88	(-.09) – (.08)	(-.05) – (.06)

* Significant at $p \leq .05$
** Significant at $p < .01$

Table 3. *Elevation Accuracy Component Analog*

Rater-Ratee Race	**Mean**		***SD***		***N***	
	Sup	Peer	Sup	Peer	Sup	Peer
Majority Rater Majority Ratee	-.10^{A}	-.09^{D}	1.29	1.25	573	1263
Minority Rater Minority Ratee	.30^{AB}	.44^{E}	1.17	1.18	60	106
Minority Rater Majority Ratee	-.01^{AB}	-.13^{D}	1.32	1.27	162	329
Majority Rater Minority Ratee	.23^{BC}	.35^{E}	1.46	1.24	149	355
Total	-.00	.00	1.32	1.26	944	2053

Note. For each rater level, means with the same letter in their superscripts do not differ significantly from one another according to a Bonferroni test with a .05 limit on family-wise error rate.

CHAPTER IV: DISCUSSION

The goal of the present study was to better explain the relative contributions of ratee performance and rater bias. The common methodology in previous meta-analytic research uses a within-ratee approach, holding ratees' constant between raters; however, this approach does not equate subgroup performance. Subsequently, it is impossible to discern from these data whether the bias resulted from majority-member supervisors understating the performance of minority-member ratees', minority-member supervisors overstating it, or a combination of both (Stauffer & Buckley, 2005). The present methodology does not equate subgroup performance, as evidenced by majority-group members performing significantly better on the standard of work performance (d = .34). However, majority- and minority-group supervisor ratings did not differ, $t(942) = .09$, $p = .93$. These true score estimates along with performance ratings were utilized to compute the components of accuracy, elevation and differential elevation, that are directly related to the accuracy of personnel decisions (Murphy et al., 1982). In turn, a rating accuracy approach is suited to help clarify both the direction and magnitude of rater-ratee race differences.

For the current study when only looking at ratings, there was no difference between group performance means from one supervisor race to the other. Majority-members did not receive higher ratings than minority-members from both minority- and majority-raters. Although neither of the following was significant: majority-ratees received higher ratings from majority-raters than from minority-raters; whereas minority-ratees received higher ratings from minority-raters than from majority-raters. The 13,862

civilian and military individuals rated by minority-raters in the Sackett & Dubois study is a far larger cumulative sample than the 528 individuals in the present study.

The current study, utilizing rating accuracy, does demonstrate a systematically larger difference between rated performance and true performance for members of one group than for another. Both minority- and majority-member raters overstated the performance of minority-member ratees. That is, both minority- and majority-member supervisors agreed on the direction and magnitude when evaluating minority-race ratees. Thus, this tendency to rate leniently is equally strong for both minority- and majority-raters. It is notable that majority- and minority-raters were just as accurate or inaccurate in their ratings for each subgroup. The practical implications of the results of this study are: If you are a majority-member ratee, both minority- and majority-member raters will accurately evaluate your performance. Therefore, it does not matter whether your supervisor is a minority- or majority-group member. If you are a minority-group ratee, you will receive an inaccurate rating overstating your performance from either a majority- or minority-group supervisor. Previous findings indicate minority workers should prefer minority supervisors (Staffer & Buckley, 2005); however, the current findings indicate it may not matter whether your supervisor is a minority- or majority-group member.

Peer raters elicited the same relationship; interestingly though, both rater subgroups significantly overstated minority ratings more than supervisor majority- or minority-member raters. This evidence leads us to conclude that both the rater level and the race of ratee do have an impact on performance ratings in real world settings. Therefore, the existence of racial bias is tenable in performance evaluations; however,

the locus of the bias is not congruent with the primary reasoning that protected demographic group members are victims of majority members understating their true performance levels (Landy, 2008).

Peer Raters Lenient

When peers evaluate minority-member ratees', both majority- and minority-raters are significantly more lenient than supervisor raters. Majority-member ratees' evaluations do not elicit this effect. This effect is opposite to what was hypothesized: peer raters were more likely to elevate (EL) ratings of peer ratees when compared with supervisor raters. Such a race effect is more likely to occur with peers than with supervisors, particularly when supervisors have received rater training (Sackett & Dubois, 1991). More importantly, what is the difference between a rater's ability and rater's motivation to provide an accurate evaluation of his or her peer. A peer should have a better opportunity to observe more work behavior; thus, resulting in sufficient observations of various job facets (Borman, 1974; Henderson, 1984; Latham & Wexel, 1982). In turn, this should lead to a more accurate judgment, as a peer rater can more easily recall critical performance episodes that have been exhibited by the ratee. Individuating information runs counter to negative stereotypes of individuals making it less likely to have minority-group members' performance understated. The different roles, orientations, and perspectives of the peer and supervisor organizational levels toward the target ratee influence peer-supervisor disagreement. For example, different rater groups may have different conceptualizations of what constitutes effective performance in a particular job (Campbell & Lee; 1988). Strong correspondence among ratings from different sources should not be expected (Lance & Woehr, 1989), but when

examined across rater level, ratings of the minority-subgroup should not be more inaccurate than the majority-subgroup.

The results of the current study suggest that negative stereotypes are not adversely affecting peer evaluations; yet, peer minority evaluations are inaccurate and more so than supervisor ratings. Peers may be in a better position and are more knowledgeable of the behaviors that are critical for favorable job performance; however, peers do not feel it incumbent on them to provide accurate ratings (Carson, Cardy, & Dobbins, 1991). It is not part of the peers' job to pay attention to and evaluate their co-workers. Moreover, peers may not attend to performance as closely, possibly depending on the interdependent nature of the peer relationship, because they have less investment in the peers' performance than do supervisors (Conway & Huffcutt, 1997). Peer raters have less accountability as they are less likely to have to justify their judgments to others (Harris, Ispas, Schmidt, 2008; Tetlock & Kim, 1987). A less substantive explanation also exists: peer raters make certain rater errors to a greater degree (e.g., leniency, restriction-in-range). That is, peer raters are in a better position for their rater capacity that could potentially insulate against the fallible social cognitive process, but they do not have the motivation to provide accurate ratings. Peer raters are more concerned with not interfering with more important tasks; damaging the co-worker relationship; demotivating their co-worker; or receiving criticism from their co-worker (Borman, 1997).

Factors Involved in the Overstating of Minority Evaluations

In theory, from a rater ability perspective, a main effect (e.g., rater race, ratee race) on performance ratings would be expected if the majority of raters in a sample

endorse or are influenced by a negative stereotype (Rudolph & Baltes, 2008). One possible factor is that raters are experiencing individuating information to weaken the negative influence of stereotypes (Landy, 2008). Considering the main effect in the current study is not in the direction congruent with the common explanation, stereotypes, should not be assumed to be affecting evaluations in the way commonly espoused by the social cognitive perspective. More importantly, there are substantial disincentives to report that any ratee is not performing well. Accurate evaluations can lead to negative repercussions for managers; and these effects could very well be exacerbated when evaluating minority-member ratees. It is likely the social cognitive process a rater is going through when rating a minority-ratee could very well be fundamentally different than when evaluating a majority-member ratee.

A substantial proportion of the population is motivated to control discrimination, as the influence of stereotypes makes individuals feel guilty and self-critical (Glaser & Knowles, 2008; Voils, Ashburn-Nardo, Monteith, & Czopp, 2002). That is, the social cognitive process in addition to the motivational mechanisms that lead to a lenient rating, such as interference with more important tasks (e.g., avoiding liability); damaging the ratee–supervisor relationship; demotivating the ratee; receiving criticism from the ratee; and, receiving criticism from the rater's supervisor (Harris, 1994) may result in an over-stated rating.

Raters are less confident and have less variability in their ratings of individuals of a different race (Schmitt & Lappin, 1980). If there is substantial variability in a ratee's performance over time; it is likely that a rater can recall critical performance episodes that are consistent with any rating that he or she chooses to give. More specifically,

individuals who either have an explicit or implicit negative common perception of minority members actively seek information that they expect to match their desired evaluation of a minority member ratee (Maynard & Brooks, 2008). When faced with deadlines, imperfect recall, or information overload (e.g., multiple subordinates and multiple situations) (Barnes-Farrell, 2001); instead of understating performance that has not been sufficiently observed, raters overstate it. When minority-member ratees have less variability in their ratings, there is a lack of attention to individual differences of performance (Greenwald, 2008). Less variability in performance evaluations do not aid in spotting individuals who perform favorable performance behaviors or possess key job-relevant skills (Tetlock, Mitchell, & Murray, 2008).

Study Limitations

There are a wide variety of issues regarding the conceptual appropriateness and empirical characteristics of various approaches to criterion measurement. The present study utilizes a walk-through performance test (WTPT) to function as a true score of participant performance (Hedge & Teachout, 1992). Subsequently, ratings of technical proficiency, technical knowledge, and the walk-through performance test were used to create Cronbach's Evaluative Accuracy Components (Cronbach, 1955). Ratings are not compared with true job performance but to another set of ratings (e.g., maximal performance) provided by a set of expert raters observing and rating performance under optimal conditions. It has been argued that there is no evidence to conclude that these expert ratings actually reflect true performance or are inherently more accurate than the other set of ratings provided by supervisors or peers (e.g., typical performance) (Woehr, 2008).

Salient to this discussion, most notably regarding the nature of Cronbach's (1955) accuracy components, is the distinction between maximal and typical measurements of performance. Essentially, maximum performance measures characterize what an individual "can do," whereas typical performance measures characterize what an individual "will do." Some individuals consistently perform at the same level, whereas other individuals' do not perform consistently (Kane, 1982). Theoretically, this distinction leads to a better understanding of the relative contribution of determinants of performance such as ability and motivation.

In the present study, the walk-through performance test (WTPT) is considered a subjective maximal performance test; while, performance ratings are also considered

subjective, their distinction as maximal or typical is not clear. According to Sackett and colleagues three conditions are necessary for a measure to be categorized as maximum performance (DuBois et al., 1993; Sackett et al., 1988) First, for the WTPT, participants knew they were being evaluated. Second, there were instructions to demonstrate their best effort. Third, measurement of the test was over a short enough period of time that a participant's best effort could be sustained. Performance ratings are theoretically more characteristic of typical performance; however, previous research has noted that performance ratings correlated more highly with maximum than typical performance. That is, in theory, performance ratings reflect performance over an extended period of time and thus should reflect typical performance; however, previous research found it may be possible that a performance rating can reflect a global judgment about ratee performance rather than a dimension-specific one. In the present study, performance ratings are considered to be typical performance measure, because they reflect performance over an extended period of time.

In the present study, Cronbach's accuracy components are consistent with the criteria for a comparison of typical and maximum performance. That is, consistent with Sackett et al. (1988) typical- and maximum-performance measures on two specific dimensions of the job, namely, technical knowledge and technical proficiency were focused on. It should be noted only two criterion dimensions were examined, this is sufficient with respect to the walk-through performance test; however, this falls short of full explication of the criterion domain for the various jobs in the sample. However, the modality of measurement is the same. Measuring typical performance subjectively (e.g., performance ratings) and maximum performance subjectively (e.g., WTPT) does not

distort the level of agreement between the two. Second, both are measured at the same level of specificity. It is appropriate to compare dimensions of typical performance (e.g., technical proficiency, technical knowledge) with a maximum performance measure of these dimensions (e.g., WTPT). That is, importantly, only the performance context changed while the content of the performance domain remained the same. Third, both measures were obtained at the same point. Finally, both typical and maximum performance measures were measured reliably. Procedures utilized in the current sample have been shown to produce accurate and reliable work-sample test ratings with high reliabilities (Hedge, Dickinson, & Bierstedt, 1988) as the administrators of the WTPT received substantial accuracy training aimed toward increasing inter-rater reliability and accurate observation of examinee performance (Hedge, Lipscomb, & Teachout, 1988).

There is support for majority-minority differences in typical versus maximal performance criteria (Dubois, Sackett, Zedeck, & Fogli, 1993). Historically, there has been a discrepancy between race differences for typical (e.g., personality measures) and maximum performance predictors (e.g., cognitive ability tests). Dubois et al. (1993) investigated whether a similar pattern might exist in typical and maximum performance criteria. They found that with respect to majority-minority differences in level of performance for typical as compared with maximum performance criteria the typical job performance of minority-members was substantially below that of majority-members. However, differences in their maximum performance were far less extreme. Their finding, although not consistent with the emphasis on ability in maximum performance predictors, is that the typical job performance of minority-members was lower than that

of majority-members, is consistent with the meta-analysis of Ford et al. (1986). The nature of typical and maximal performance in the context of majority-minority differences has only been investigated in a narrow sense (i.e., cashiers). The current study's findings are not consistent with the findings of Dubois et al., (1993) and Ford et al., (1986). Minority-member ratings of typical performance are greater than their corresponding maximal performance measurement. One could potentially interpret that minority-individuals perform at a higher level on a day-to-day basis, as rated by their supervisors and peers, when compared to their maximal performance assessment. This explanation does not explain the influence of rater-ratee race compositions role of affecting the level of agreement both within and between ratings sources. As noted previously, under affirmative action and equal employment pressures, race is likely to be an important consideration potentially motivating supervisors to rate leniently with respect to minority-group members to reduce the likelihood of legal scrutiny associated with low ratings (Kraiger & Ford, 1985). This concern provides a direction for future research.

Another study limitation is the amount of power of the study when compared to other research investigating racial bias. Within each rater level, ratees in the current sample were not rated by both majority- and minority-member raters. Therefore, the present study does not utilize a repeated measures methodology. A repeated measures methodology, consistent with Sackett & Dubois or Stauffer & Buckley, in addition to having more power afforded by a much larger sample size, has more power than an individual samples methodology to detect significant effects. With additional power, marginal effects in the current sample could potentially become significant.

Furthermore, when computing the differential elevation component, many rater-ratee race cell means had a very small sample size.

Directions for Future Research and Practice

Both the research and practice of performance evaluations are integrally concerned with answering the question of what a rater is trying to do when he or she completes a performance rating form. Too much emphasis is being placed on a rater's capability to use their best judgment than on their willingness to provide honest evaluations of ratee performance (Murphy, 2008). Not enough attention is given to the role the rater plays in the organizational context of ratings in high stakes organizational decisions. That is, in organizations there are real and tangible consequences to unfair or biased evaluations (Landy, 2008). It could be that in the current racial climate of affirmative action and equal employment pressures, raters may be motivated to not give a minority-member an unfavorable rating. Importantly, conditions have not yet been established in which negative stereotypes are likely to operate in an organizational context (Maynard & Brooks, 2008). Boundary conditions need to be delineated for when raters are trying their best to accurately measure the performance of their subordinate or peers; and when, they are motivated to not provide accurate measures to the best of their ability (Banks & Murphy, 1985; Harris et al. 2008; Hollenbeck 2008; King 2008). Future research need not disentangle the influences concerning rater ability and rater motivation, but rather investigate how a rater’s ability is affected by his or her motivation to avoid undesirable consequences from giving unfavorable ratings.

In the present study, stereotypes of raters were not measured, yet given the results it does not make sense to argue that many individuals have negative

stereotypes that are adversely affecting evaluations of minorities. The present study suggests motivational factors of performance evaluation are probably a more salient influence on ratings. That is, if the social cognitive theory posits that raters have a tendency to associate the minority group ratee individuals with their low performance schema and are less likely to associate these individuals with their high performance schema, then how could it be that raters are over-stating performance of minority-member ratees?

Future research needs to investigate rater's perceptions and attitudes when rating minority-member ratees. It is possible that negative stereotypes regarding racial minorities could engender raters to over-correct, thus inflating ratings of minority members (Tetlock, Mitchell, & Murray, 2008). In this way ratings are influenced by the raters' reactance toward negative stereotypes and discomfort with giving negative feedback, rather than an assessment of the ratees' actual performance behavior. If a rater does not convey his honest impression – his or her evaluation is not consistent with his or her honest assessment of an individual, how will the mechanisms of social cognition be affected? It is likely the rater will compensate by overly attending to, encoding, storing, and recalling incidents of satisfactory performance, however rare, that corroborate the rating (e.g., cognitive dissonance). Furthermore, future research needs to assess individual differences as to who is more likely to tolerate the disparity between their real impressions of a ratee and the inflated rating he or she may be motivated to elicit.

Stereotyping may not adversely affect evaluations but can affect the performance of commonly stereotyped groups; that is, raters engage in social interaction, often in a

manner that reinforces their initial low expectations (Heilman & Eagly, 2008) – while, minority-ratees reciprocate by engaging in self-limiting behaviors (Kraiger & Ford, 1985). This may not lead to a negative evaluation; but possibly more importantly, when assessing the job performance of minority-members, these individuals may not be receiving appropriate feedback. Future research needs to investigate how inaccurate ratings influence the matriculation of an employee throughout the organization. There is evidence that accurate feedback increases self-awareness and performance in weak areas (e.g., Atwater et al., 1995). That is, receiving performance information that is highly correlated with actual ratee performance levels is important to properly guide behavioral change. More specifically, if changes in behavior are needed to improve performance, then ratees must be made aware through precise and focused knowledge about the areas needing improvement (Borman, 1997). Performance evaluations are often tied to compensation, promotion, and positive standing in regards to downsizing. When evaluations are made for the allocation of rewards, it is reasonable to assume that a favorable rating will lead to better outcomes in the short term; however, there are no data to support the idea that there is a significant relationship between initial ratings and the ascendance to top management positions. It can be assumed that early career mobility will lead to eventual success, but the role of performance evaluations in long-term career success is uncertain (Landy, 2008). It could be that an individual's opportunity to receive performance information that is highly correlated with his or her actual performance level will better predict the likelihood of his or her advancement to top management positions. This may be an explanation why there is a lack of mentors for minorities in top management positions. This lack of mentorship opportunities

coupled with insufficient performance feedback contribute to the disadvantage of minority-group members (Ilgen & Youtz, 1984). Future research efforts should be directed at better understanding the nature of the disparity between ratees of different races, particularly the inaccuracy when assessing the job performance of minority-member ratees.

Race an Important Consideration in Reliability and Validity of Multisource Feedback

The results of the current study indicate the rater-ratee race composition can potentially affect the level of agreement both within and between ratings sources. In theory, lack of agreement across sources may reflect true differences resulting from differences in perspectives or opportunities to observe performance (Woehr, 2008). While across-source disagreement signals a lack of reliability it may not signal a lack of validity. It is important to note high within-source agreement is desired and suggests that the rating source is providing a valid view of performance (Borman, 1997). That is, multiple raters are essential to increase reliability (Ones, Viswesvaran, Schmidt, 2008), as within-source agreement can be increased by obtaining larger numbers of raters (Borman, 1997). Interestingly, in the current study both rater subgroups agree on each subgroup, therefore, reliability may not be affected. In general, ratings from differently situated raters such as peers and supervisors can be justifiably pooled for decision making (Ones, Viswesvaran & Schmidt, 2008); however, the current study indicates there is a difference across rating sources when evaluating minority members.

Future research needs to investigate how the rater-ratee race composition affects ratings from more than one organizational level and if this contextual factor

affects the validity of rating information if there is disagreement across rating levels. Subsequently this would help determine if certain rater-ratee race combinations, which are more likely to have both within-level and across-level disagreement also are not as valid. It should be noted when minority-member ratees have less variability in their ratings, there is a lack of attention to individual differences in performance (Greenwald, 2008). Less variability in performance evaluations does not aid in spotting individuals who perform favorable performance behaviors or possess key job-relevant skills (Tetlock, Mitchell, & Murray, 2008); subsequently, validity of ratings would be attenuated.

Conclusion

Ratee race and rater level contribute to differences in the accuracy of ratings. From a rater perspective, ratees are being differentially evaluated across subgroups. From an individual ratee perspective, a rater from the majority- or minority-group will give the same rating. Peer raters are more likely to be more lenient with minority-ratees than with majority-member ratees. How are raters arriving at these differences? Accuracy of these ratings is paramount to identifying strengths and weaknesses of workers and may influence an individual's ascension through an organization. Further research needs to address how race affects reliability within-rater level and validity of ratings over time.

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*, 203-231.

Barnes-Farrell, J. (2001). Performance appraisal: Person perception processes and challenges. In M. London (Ed.), How people evaluate others in organizations (pp. 135–153). Mahwah, NJ: Lawrence Erlbaum.

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

Bernardin, H. J., & Beatty, R. W. (1984). Performance appraisal: Assessing human behavior at work. Boston: Kent.

Bielby, W. T. (2000). Minimizing workplace gender and racial bias. *Contemporary Sociology*, 29,*1*, 120–129.

Borman, W. C. (1974). The rating of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance*, *12*, 105–124.

Borman, W. C. (1997). 360° ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review*, *7*, 299–315.

Cardy, R.L., & Dobbins, G.H. (1994). Performance Appraisal: Alternative Perspectives. Cincinnati, Ohio: South-Western Publishing Co.

Cardy, R. L., & Krzystofiak, F. J. (1991). Interfacing high technology operations with blue collar works: Selection and appraisal in a computerized manufacturing setting. *Journal of High Technology Management Research*, *2*, 193-210.

Carroll, S. J., & Schneir, C. E. (1982). Performance appraisal and review systems: The identification, measurement, and development of performance in organizations. Glenview, IL: Scott, Foresman.

Carson, K. P., Cardy, R. L. & Dobbins, G. H. (1991). Performance appraisal as effective management or deadly management disease: Two empirical investigations. *Group and Organization Studies*, *16*, 143-159.

Christal, R. E. (1974). The United States Air Force occupational research project (AFHRL-TR-73–75). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Occupational Research Division.

Copus, D. (2005). A lawyer's view: Avoiding junk science. In F. J. Landy (Ed.), Employment discrimination litigation: Behavioral, quantitative, and legal perspectives (pp. 450–462). San Francisco: Jossey-Bass.

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

Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1988). The effects of purpose of appraisal and individual differences in stereotypes of women on sex

differences in performance ratings: A laboratory and field study. *Journal of Applied Psychology*, *73*, 551-558.

DuBois, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and white-black differences. *Journal of Applied Psychology*, *78*, 205-211.

Eagly, A. (1995). The science and politics of comparing women and men. *American Psychologist*, *50*, 145-158.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology*, *86*, 215–227.

Fayol, H. (1949). General and industrial management. New York: Pitman.

Feldman, J. M. (1986). Instrumentation and training for performance appraisal: A perceptual-cognitive viewpoint. In G. R. Ferris & K. M. Rowland, Eds., Research in personnel and human resources management, vol. 4. Greenwich, CT: JAI Press.

Ferris, G. R., & Judge, T. A. (1991). Personnel/human resources management: A political influence perspective. *Journal of Management*, *17*, 1-42.

Ford, J. K., Kraiger, K., & Schechtman, S. L. (1986). Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. *Psychological Bulletin*, *99*, 330–337.

Funder D. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*, 75-90.

Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology*, *44*, 164–172.

Greenwald, A.G. (2008). Landy is correct: Stereotyping can be moderated by individuating the out-group and by being accountable. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 430–435.

Harris, M. M., Ispas, D. and Schmidt, G. F. (2008), Inaccurate performance ratings are a reflection of larger organizational issues. *Industrial and Organizational Psychology*, 1, 190–193. doi: 10.1111/j.1754-9434.2008.00037.x

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self–supervisor, self–peer, and peer–supervisor ratings. *Personnel Psychology*, *41*, 43–62.

Hastie, R. (1981). Schematic principles in human memory. In E. T. Higgins, C. P. Hrman, & M. P. Zanna, Eds., *Social cognition: The Ontario symposium*, vol. 1. Hillsdale, NJ: Erlbaum. Pp. 39-88.

Hedge, J. W., Dickinson, T. L., & Bierstedt, S. A. (1988). The use of videotape technology to train administrators of Walk-Through Performance Testing (AFHRL-TP-87–71). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Training Systems Division.

Hedge, J. W., Lipscomb, M. S., & Teachout, M. S. (1988). Work sample testing in the Air Force job performance measurement project. In M. S.Lipscomb & J. W.Hedge (Eds.), Job performance measurement: Topics in the performance measurement of Air Force enlisted personnel (AFHRL-TP-87–58). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Training Systems Division.

Hedge, J. W., & Teachout, M. S. (1986). Job performance measurement: A systematic program of research and development (AFHRL-TP-86–37). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Training Systems Division.

Hedge, J. W., & Teachout, M. S. (1992). An interview approach to work sample criterion development. *Journal of Applied Psychology*, *77*, 453–462.

Ilgen, D. R., & Youtz, M. (1984, February). Factors affecting the evaluation and development of minorities in organizations Paper presented at the Office of Naval Research Conference on Minorities Entering High-Technology Careers, Pensacola, FL.

Ivancevich, J. M. (2001). *Human resource management*. New York, NY.

Johnson, J. W., & Cochran, C. C. (2008). Studying the influence of stereotypes on personnel decisions in the real world. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 423–425.

Kane, J. S. (1982, November). *Rethinking the problem of measuring performance: Some new conclusions and a new appraisal method to fit*

them. Paper presented at the 4th Johns Hopkins University National Symposium on Educational Research, Washington, DC.

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

Klimoski, R. J., & London, M. (1974). Role of the rater in performance appraisal. *Journal of Applied Psychology*, *59*, 445–451.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*, *70*, 56–65.

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, *129*, 522–544.

Landy, F. J. (2008). Stereotypes, bias, and personnel decisions: Strange and stranger. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 379–392.

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

Latham, G. P., & Wexley, K. N. (1982). Training approaches and a workshop to minimize rating error. In L. S. Baird, R. W. Beatty, & E. E. Schneir, Eds., The performance appraisal sourcebook. Amherts, MA: Human Resource Development Press, pp. 85-90

Lord, R. G. and Maher, K. J., 1991. Cognitive theory in industrial and organizational psychology. In: Dunnette, M.D. and Hough, L.M., Editors, 1991. *Handbook of industrial and organizational psychology,* Consulting Psychologists Press, Palo Alto, CA.

Maass, A., Castelli, L., & Arcuri, L. (2000). Measuring prejudice: Implicit versus explicit techniques. In D.Capozza & R.Brown (Eds.), *Social identity processes: Trends in theory and research* (pp. 96–116). London: Sage.

Martell, R. F., Lane, D. M., & Emrich, C. (1996). Male-female differences: A computer simulation. *American Psychologist*, *51*, 157–158.

Maynard, D. C., & Brooks, M. E. (2008). The persistence of stereotypes in the context of familiarity. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 417–419.

McKay P. F., & McDaniel M.A. (2006). A reexamination of Black-White mean differences in work performance: More data, more moderators. *Journal of Applied Psychology*, *91*, 538-554.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002).Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology*, *83*, 1029–1050.

Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 148–160.

Murphy K. R., Balzer W.K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*, 619-624.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

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

Neisser, U. (1967). *Cognitive psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Nkomo, S. M. & Cox, T. H., Jr. (1989). Gender differences in the upward mobility of black managers: Double whammy or double advantage? *Sex Roles*, *21*, 825-839.

Norman, C. A., & Zawacki, R. A. (1991, December). Team appraisals–team approach. *Quality Digest*, *11*, 68–75.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2008). No new terrain: Reliability and construct validity of job performance ratings. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 174–179.

Osterman, P. (2007). Comment on Le, Oh, Shaffer and Schmidt. *Academy of Management Perspectives*, *3*, 16–18.

Peterson N. G., Borman W. C., Hanson M. A., & Kubisiak U.C. (1999). Summary of results, implications for O*NET applications, and future directions. In

Peterson NG, Mumford MD, Borman WC, Jeanneret PR, Fleishman EA (Eds.), An occupational information system for the 21st century: The development of O*NET. Washington, DC: American Psychological Association.

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology*, *74*, 770–780.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, *35*, 1-62.

Roach, D. W., & Gupta, N. (1992). A realistic simulation for assessing the relationships among components of accuracy. *Journal of Applied Psychology*, *77*, 196–200.

Rosenbaum, J. E. (1979). Tournament mobility: Career patterns in a corporation. *Administrative Science Quarterly*, *24*, 220-241.

Rotundo M., Sackett P. R. (1999). Effect of rater race on conclusions regarding differential prediction in cognitive ability tests. *Journal of Applied Psychology*, *84*, 815–822.

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

Sackett, P. R., & DuBois, C. L. Z. (1991). Rater-ratee race effects on performance evaluation: Challenging meta-analytic conclusions. *Journal of Applied Psychology*, *76*, 873–877.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, *73*, 482-486.

Schmidt, F. L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior*, *33*, 272–292.

Shore T. H., Shore L. M., Thornton G. C. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. Journal of Applied Psychology, *77*, 42–54.

Stauffer, J. M., Buckley, M. R. (2005). The existence and nature of racial bias in supervisory ratings. *Journal of Applied Psychology*, *90*, 586–591.

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology*, *52*, 700-709.

Tetlock, P. E., Mitchell, G., & Murray, T. L. (2008). The challenge of debiasing personnel decisions: Avoiding both under- and overcorrection. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 439–443.

Wessel, J. L., & Ryan, A. M. (2008). Past the first encounter: The role of stereotypes. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 409–411.

Wigdor, A., & Green, B. F., Jr. (1991). *Performance assessment for the workplace* (Vols. 1 and 2). Washington, DC: National Academy Press.

Woehr, D. J. (2008). On the relationship between job performance and ratings of job performance: What do we really know? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 161–166.

Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, *10*, 109–120.

Appendix A: IRB Documentation

University and Medical Center Institutional Review Board
East Carolina University • Brody School of Medicine
600 Moye Boulevard • Old Health Sciences Library, Room 1L-09 • Greenville, NC 27834
Office 252-744-2914 • Fax 252-744-2284 • www.ecu.edu/irb
Chair and Director of Biomedical IRB: L. Wiley Nifong, MD
Chair and Director of Behavioral and Social Science IRB: Susan L. McCammon, PhD

TO: Mark C. Bowler, Dept of Psychology, ECU—104 Rawl Building

FROM UMCIRB

DATE: May 22, 2009

RE: Human Research Activities Determined to Meet Exempt Criteria

TITLE: "Demographic Influences on Performance Appraisal Ratings"

UMCIRB #09-0472

This research study has undergone IRB review on 5.20.09. It is the determination of the IRB Chairperson (or designee) that these activities meet the criteria set forth in the federal regulations for exemption from 45 CFR 46 Subpart A. These human research activities meet the criteria for an exempt status because it is a research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. *NOTE: 1) This information must be existing on the date this IRB application is submitted. 2) The data collection tool may not have an identifier or code that links data to the source of the information.*

. The Chairperson (or designee) deemed this **unfunded** study **no more than minimal risk.** This research study does not require any additional interaction with the UMCIRB unless there are proposed changes to this study. Any changes must be submitted to the UMCIRB for review prior to implementation to allow determination that proposed changes do not impact the activities eligibility for exempt status. Should it found that a proposed change does require more substantive review, you will be notified in writing within five business days.

The following items were reviewed in determination exempt certification:

- Internal Processing Form – Exempt Application
- Letter of Support

It was furthermore determined that the reviewer does not have a potential for conflict of interest on this study.

The UMCIRB applies 45 CFR 46, Subparts A-D, to all research reviewed by the UMCIRB regardless of the funding source. 21 CFR 50 and 21 CFR 56 are applied to all research studies that fall under the purview of Food and Drug Administration regulations. The UMCIRB follows applicable International Conference on Harmonisation Good Clinical Practice guidelines.