Differential Prediction: Understanding a Tool for Detecting Rating Bias in Performance Ratings

by

Emilee B. Tison

Thesis submitted to the Faculty of
Virginia Polytechnic Institute and State University
In partial fulfillment of the requirements for the degree of

Master of Science
In
Psychology

Neil M.A. Hauenstein, Ph.D., Chair
Roseanne J. Foti, Ph.D.
Robert S. Stevens, Ph.D.

March 19th, 2008
Blacksburg, VA 24060

Keywords: Differential Prediction, Performance Ratings, Race Effects, Rating Bias,
Intercept Bias
DIFFERENTIAL PREDICTION: UNDERSTANDING A TOOL FOR DETECTING RATING BIAS IN PERFORMANCE RATINGS

By

Emilee B. Tison

Neil M.A. Hauenstein, Chairperson

Psychology

(ABSTRACT)

Three common methods have been used to assess the existence of rating bias in performance ratings: the total association approach, the differential constructs approach and the direct effects approach. One purpose of this study was to examine how the direct effects approach, and more specifically differential prediction analysis, is more useful than the other two approaches in examining the existence of rating bias. However, the usefulness of differential prediction depends on modeling the full rater race X ratee race interaction. Therefore, the second purpose of this study was to examine the conditions where differential prediction has sufficient power to detect this interaction. This was accomplished using monte carlo simulations. Total sample size, magnitude of rating bias, validity of predictor scores, rater race proportion and ratee race proportion were manipulated to identify which conditions of these parameters provided acceptable power to detect the rater race X ratee race interaction; in the conditions where power levels are acceptable, differential prediction is a useful tool in examining the existence of rating bias. The simulation results suggest that total sample size, magnitude of rating bias and rater race proportion have the most impact on power levels. Furthermore, these three parameters interact to effect power. Implications of these results are discussed.
Acknowledgements

I would like to thank my advisor, Dr. Neil Hauenstein, for his guidance and continued support on this project. It is because of his tremendous amount of knowledge, patience and understanding that this project is now complete; I am eternally grateful to have the opportunity to learn from him. His enthusiasm sparks my own, and drives me to persevere.

I would also like to thank the other members of my committee for their feedback and valuable insights: Dr. Roseanne Foti and Dr. Robert Stephens. It has been an honor to work with each and every one of them.

This project would probably still be ongoing had it not been for the help of Shannon Fraker. She possessed skills that I did not have and graciously shared them with me. For that, I am thankful.

I wish to acknowledge my family. They take interest in all that I do and provide support, as well as words of wisdom, along the way. This venture would have been much more difficult without them.

It is also important that I acknowledge a couple of my fellow graduate students: Tanner Bateman, Jaron Holmes and Matthew Jarrett. They have been a tremendous asset, acting as sounding boards for my ideas and keeping me focused on the task at hand. More importantly, they made this entire process a lot more fun. I thank them and wish them the best as they continue on.

Finally, I would like to thank my husband, Scott. He is a source of continual encouragement and unconditional love. He never fails to provide me with the strength and
confidence to continue on with my endeavors; without him, I would not have made it this far.

This project is dedicated to him.
Table of Contents

Abstract ................................................................................................................................. ii
Acknowledgments ................................................................................................................ iii
Introduction .......................................................................................................................... 1
  Detecting Rating Bias ........................................................................................................ 3
    Total Association Approach ......................................................................................... 4
      Key Total Association Studies .................................................................................. 4
    Summary ....................................................................................................................... 7
  Differential Constructs .................................................................................................. 9
    Key Differential Constructs Studies .......................................................................... 9
    Summary ....................................................................................................................... 12
  Direct Effects Approach .............................................................................................. 14
    Key Direct Effects Study ............................................................................................ 15
    Differential Prediction and the Direct Effects Approach ......................................... 16
    Key Differential Prediction Studies .......................................................................... 17
    Summary ....................................................................................................................... 21
Overview ........................................................................................................................... 21
  Purpose ............................................................................................................................ 22
Methods ............................................................................................................................ 23
  Simulation Program ....................................................................................................... 23
  Manipulated Parameters ............................................................................................... 24
    Validity Coefficient ...................................................................................................... 25
    Magnitude of Rating Bias ............................................................................................ 25
    Total Sample Size ....................................................................................................... 25
    Rater Race Proportion .................................................................................................. 25
    Ratee Race Proportion ................................................................................................. 25
Analyses ............................................................................................................................. 26
Results.......................................................................................................................26

Power Analysis.......................................................................................................26
  Small Rating Bias.............................................................................................27
  Moderate Rating Bias......................................................................................27
  Large Rating Bias............................................................................................28

Summary...............................................................................................................29

Components of Variance Analyses....................................................................29
  Main Effects....................................................................................................29
  Interactions.....................................................................................................30

Discussion.............................................................................................................32
  Main Effects....................................................................................................33
  Interactions.....................................................................................................36
  Summary.........................................................................................................38
  Conclusions....................................................................................................39

References...........................................................................................................41

Tables..................................................................................................................45

Figure Captions..................................................................................................52

Figures................................................................................................................54
List of Tables

Table 1: Proportion of variance accounted for in power levels by the manipulated parameters……………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………
List of Figures

Figure 1: Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.30, validity = 0.10………………………….54

Figure 2: Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.30, validity = 0.30………………………….55

Figure 3: Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.30, validity = 0.50………………………….56

Figure 4: Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.30, validity = 0.10………………………….57

Figure 5: Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.30, validity = 0.30………………………….58

Figure 6: Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.30, validity = 0.50………………………….59

Figure 7: Power results for rater race proportion and ratee race proportion combinations: total sample size = 500, magnitude of bias = 0.45, validity = 0.10………………………….60

Figure 8: Power results for rater race proportion and ratee race proportion combinations: total sample size = 500, magnitude of bias = 0.45, validity = 0.30………………………….61

Figure 9: Power results for rater race proportion and ratee race proportion combinations: total sample size = 500, magnitude of bias = 0.45, validity = 0.50………………………….62
Figure 10: Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.45, validity = 0.10………………………….63

Figure 11: Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.45, validity = 0.30………………………….64

Figure 12: Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.45, validity = 0.50………………………….65

Figure 13: Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.45, validity = 0.10………………………….66

Figure 14: Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.45, validity = 0.30………………………….67

Figure 15: Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.45, validity = 0.50………………………….68
Introduction

Performance ratings are the most commonly used method of performance appraisals (e.g. Landy & Farr, 1980; Landy & Shankster, 1994). Measurements of performance, and more specifically performance ratings, have assisted personnel decisions from providing information for hiring, feedback development and indices of employee effectiveness to working as criteria for validation studies and training evaluations (cf., Arvey, 1979; Landy & Farr, 1980; Latham & Wexley, 1981). Performance ratings are also used in various administrative decisions such as promotions, raises and terminations (Cardy, 1998). In short, many personnel decisions are made with the assistance of performance ratings. When making personnel decisions, however, the goal is to ensure that the methods employed remain unbiased. Bias refers to “any construct-irrelevant source of variance that results in systematically higher or lower scores for identifiable groups of examinees (p. 30, Society for Industrial and Organizational Psychology (SIOP), 2003).” In other words, systematic differences between groups exist that are not attributable to the construct being rated. Therefore, if performance ratings are biased, the decisions based off of them would be unwarranted. Many researchers have expressed concern that the assumption of unbiased performance ratings could be unfounded (Rotundo & Sackett, 1999). This raises an important question: are performance ratings unbiased measures of performance, i.e., does rating bias exist in performance ratings?

Landy and Farr (1980) attempted to draw conclusions regarding performance ratings by reviewing the literature up to 1980. In doing so, they addressed potential sources of variation in performance ratings that included Roles, Vehicle and Context (see Landy & Farr, 1980, for a more in-depth discussion involving the sources of variation in performance ratings). Roles referred to rater characteristics, ratee characteristics and the interaction between rater and ratee
characteristics. This included such characteristics as race, gender, age, and experience, among others (cf. Landy & Farr, 1980). Vehicle represented rating formats (e.g., direct rating and derived rating systems) and other technical issues such as rating dimensions, the number of response categories, and anchors. Context signified the context in which the rating occurred, “not explicitly related to the nature of the rater, ratee or rating instrument (p. 89, Landy & Farr, 1980).” In other words, context included such things as the intended use of the ratings, position or job characteristics, and supervisor leadership style. It is evident that there are many potential sources of variation in performance ratings; however, this study focuses on Roles. More specifically, this study concentrates on the rater and ratee characteristic of race.

Even though “few general conclusions” were made regarding the effects of rater characteristics (i.e., race) on performance ratings (p. 77) in Landy and Farr (1980), more distinct trends were uncovered regarding rater race X ratee race interactions; raters tended to give more favorable ratings to ratees of the same race (Landy & Farr, 1980). Despite the tendency of same race bias, Landy and Farr (1980) contended that “no direct results are available that bear on this question” (p. 82). The potential of same race bias in performance ratings did not gain strength until Kraiger and Ford’s (1985) meta-analysis. Kraiger and Ford (1985) emphasized the potential of same race rating bias in performance ratings and, in doing so, generated interest examining the existence of rating bias in performance ratings (e.g. Martocchio & Whitener, 1992; Pulakos, White, Oppler & Borman, 1989; Rotundo & Sackett, 1999; Sackett & DuBois, 1991). For example, Pulakos, White, Oppler and Borman (1989) criticized Kraiger and Ford’s (1985) study and concluded that any evidence of racial bias in performance ratings account for too little variance to be of any importance. Others mirrored this sentiment and the research following Kraiger and Ford (1985) suggested performance ratings lack evidence of rating bias
(e.g. Landy & Shankster, 1994; Latham & Wexley, 1981; Waldman & Avolio, 1991). More recent research, however, has come full circle regarding conclusions of rating bias and researchers are concluding the existence of rating bias in performance ratings (e.g. Stauffer & Buckley, 2005; Hauenstein & Tison, 2007).

In regards to the above research, most of the empirical evidence of the presence or absence of rating bias was based on the comparison of ratee race subgroup means as a function of rater race. Although evaluating ratee subgroup means as a function of rater race is a useful technique for detecting rating bias, there is one key limitation: rating bias cannot be separated from true ratee race performance differences. The first purpose of the current study is to build on the argument that differential prediction analyses provide a better method for detecting rating bias by providing a strategy of separating rating bias from true performance differences (Rotundo & Sackett, 1999). The second purpose of the study is to establish the boundaries in regards to the technical feasibility of using differential prediction analyses to detect rating bias. More specifically, I will conduct simulations to determine the conditions that provide adequate power for the detection a significant aptitude by ratee race X rater race interaction.

Detecting Rating Bias

Although the comparison of subgroup mean differences is the most common strategy of detecting rating bias, there are other choices. In fact, there are three “commonly used methods” for assessing bias in performance measures (Oppler, Campbell, Pulakos, & Borman, 1992): the total association approach, the differential constructs approach and the direct effects approach.
Total Association Approach

The total association approach attempts to determine the amount of criterion variance that is attributable to subgroup membership (Oppler et al, 1992), i.e. comparison of subgroup means. This subgroup membership refers not only to the subgroup of the ratee, but also to the subgroup of the rater and the interaction between the subgroup of the ratee and the subgroup of the rater (Mount et al, 1997).

The examination of the impact of rater and ratee characteristics on performance ratings is commonly conducted in the literature (Landy & Farr, 1980; Landy & Shankster, 1994) as is the total association approach of investigating the rater race main effects, ratee race main effects and rater race by ratee race interaction effects on performance ratings (e.g. Kraiger & Ford, 1985; Oppler, Campbell, Pulakos & Borman, 1992, Stauffer & Buckley, 2005). Beyond the issue of the confound between rating bias and true performance differences, sampling error is also a concern when comparing subgroup means. For between-ratee designs, where each ratee is evaluated by either a White or Black rater, large sample sizes are the best strategy to alleviate concerns about sampling error. On the other hand, within-ratee designs, where each ratee is evaluated by both a White and Black rater, eliminates this sampling error and is the best method for detecting rating bias when only using subgroup means.

Key total association studies. The most commonly cited meta-analysis that examines race effects in performance ratings is Kraiger and Ford (1985). Kraiger and Ford (1985) obtained 74 studies (30 published and 44 unpublished) for their meta-analysis in which both White and Black raters rated White and Black ratees. Examination of their estimated $d$ statistics indicated that White raters tended to rate White ratees higher than Black ratees and Black raters
tended to rate Black ratees higher than White ratees. They concluded that there exists a same race effect in performance ratings, i.e. a rater race by ratee race interaction. However, Kraiger and Ford (1985) conceded that assertions regarding bias are difficult given the fact that this data does “not directly isolate the effects of racial bias or performance differences in performance evaluations (p. 61).”

Pulakos, White, Oppler and Borman (1989), using data from the United States Army’s Project A, conducted between-subject analyses on more than 39,000 rater-ratee pairs and within-group analyses on more than 1800 ratees. Pulakos et al (1989) found significant race effects, both in the between- and within-group analyses. However, they fail to conclude rating bias for two reasons: 1) with such a large sample size, i.e. 39,000 rater-ratee pairs, even “trivial” differences were statistically significant and 2) the race effects accounted for minimal amounts of variance between- and within-ratees. Therefore, they concluded that despite the statistical evidence of rating bias, the practical consequences of rating bias were minimal (Pulakos et al, 1989). Oppler et al. (1992) essentially replicated the same analyses as Pulakos et al (1989), and produced the same empirical findings.

Along the same lines, Sackett and DuBois (1991) challenged the argument that raters rate ratees of the same race more favorably than other ratees (Kraiger & Ford, 1985). Using civilian data from the U.S. Employment Service’s General Aptitude Test Batter (GATB) and military data from the Army’s Project A, Sackett and DuBois (1991) conducted both between- and within-subject ANOVA analyses. The between-subject design consisted of 19,765 rater-ratee race pairs (12,022 White rater – White ratee, 661 Black rater – White ratee, 5,972 White rater – Black ratee, and 1,110 Black rater – Black ratee pairs) and the within-subject design consisted of 286 White ratees and 331 Black ratees. Examining the performance rating means, the results
indicated that Black ratees were rated lower than White ratees, for both White and Black raters. In other words, Black raters did not rate Black ratees higher than they rate White ratees. This pattern of findings was similar across both the between- and within-group analyses. Sackett and DuBois (1991) conclude that evidence for the same race bias (i.e. raters rate more favorably ratees of the same race) was not found. Rather, both White and Black raters rated Black ratees lower than White ratees, and they suggest that the differences may reflect true performance differences between Black and White ratees.

Mount, Sytsma, Hazucha and Holt (1997), examined the rating bias issue within one organization. “The use of a single occupational group reduces the potential compounding due to differences in job content (p. 56, Mount et al, 1997).” Using 360° feedback, i.e. supervisor, peer and subordinate ratings, Mount et al (1997) conducted both between- and within-subjects analyses. The between-group analyses were based on an average of 39,464 ratings of White managers and 1019 ratings of Black managers across the three rater levels, i.e. supervisor, peer and subordinate ratings, while the within-group analyses were based on an average of 468 White managers and 68 Black managers across the three rater levels. Overall, the results from these analyses were similar to Pulakos et al (1989) in the sense that consistent rater race by ratee race effects were detected. However, Mount et al (1997) conclude that race effects are present, citing that “Black raters assign higher ratings than Whites [raters] (p. 62)” and furthermore, “Blacks rate Blacks higher than Whites [do] (p. 68).” Mount et al (1997) suggest that their results support the existence of rating bias because White and Black raters differ on the effect size of performance ratings for White and Black ratees. The magnitude of the difference in performance ratings of Black and White ratees differ across Black and White raters, where Black raters rate Black ratees significantly higher than do White raters. These results, however, stem from
developmental ratings which were used to provide feedback to managers. Therefore, any
evidence of rating bias could be due to the developmental purpose of the ratings and potentially
limit the generalizability of these results (Mount et al, 1997).

Stauffer and Buckley (2005) re-analyzed the data from Sackett and DuBois (1991) (i.e.,
Recall that Sackett and DuBois (1991) were interested in detecting rating bias in the form of
same race bias, i.e., whether raters rated ratees of the same race higher than ratees of a different
race. Stauffer and Buckley (2005), however, were interested in whether White and Black raters
agreed on the magnitude of performance rating differences between White and Black ratees.
Therefore, they tested for the rater race X ratee race interaction. Such an interaction suggests
that “the two groups of supervisors (Black and White) do not agree on the magnitude of any
performance difference between the two groups of workers (p. 588, Stauffer & Buckley, 2005).”
Stauffer and Buckley (2005) reported that both the White and Black raters favored White ratees
over Black ratees. The White and Black raters did not agree, however, on the magnitude of the
difference between ratings of White ratees and the ratings of Black ratees. White and Black
raters differ 0.30 SD (in the examination of the civilian data) in mean job performance ratings for
White and Black ratees (Stauffer & Buckley, 2005). In other words, Stauffer and Buckley
(2005) provided evidence for a significant rater race X ratee race interaction and concluded that
rating bias exists in performance ratings. Consequently, Stauffer and Buckley (2005) suggest
continued research on rating bias.

Summary. The examination of rating bias, through the total association approach, has
produced similar empirical findings. Research commonly present evidence where both White
and Black raters rate White ratees higher than Black ratees. Furthermore, the ratee race effect is
larger for White raters than Black raters. The key issue is that researchers commonly produce evidence for this interaction effect where Black raters tend to rate Black ratees higher than do White raters. Despite this consistent pattern of findings, the interpretations differ greatly. Pulakos et al (1989) concluded that the rater race X ratee race interaction effect accounted for too little variance within-groups to be considered practically significant. Likewise, Sackett and DuBois (1991) concluded a lack of rating bias in performance ratings, but suggested this was due to the agreement between White and Black raters to rate White raters higher than Black raters. Mount et al (1997), on the other hand, suggested that rating bias is present in performance ratings and presented evidence that White and Black raters differ in their magnitude of differences in mean performance rating scores for White and Black ratees (i.e. a rater race by ratee race interaction effect). Similarly, Stauffer and Buckly (2005) found that White and Black raters differ 0.30 SD in their ratings of Black ratees. Concluding the existence of rating bias in performance ratings, Stauffer and Buckly (2005) suggested that not only does the direction of performance rating across White and Black raters need to match, but the magnitude of their differences in performance ratings for White and Black ratees must also match before the assertion of no rating bias is made.

It is interesting that findings using the total association approach are very similar, yet the interpretations of these consistent findings vary. These differing interpretations are the manifestation of the confound between true ratee race effects on job performance and rating bias. This is the fundamental problem of the total association approach; true ratee race performance differences cannot be separated from rating bias (Oppler et al, 1992).
**Differential Constructs**

The differential constructs approach attempts to examine the relationship between performance ratings and other criterion measures (Oppler et al, 1992), such as objective measures of job performance, to aid in the interpretation of performance ratings across the ratees, the raters, or both (Oppler et al, 1992; Mount et al, 1997). In other words, the goal of differential constructs approach is to control for true performance differences across ratees by examining the convergent validity of performance ratings with other variables, such as objective measures of job performance, that are assumed to be unbiased. In doing so, researchers can examine how performance ratings differ as a function of rater race, ratee race and the rater race X ratee race interaction.

*Key differential construct studies.* In a follow-up to the Kraiger and Ford (1985) study, Ford, Kraiger and Schechtman (1986) included 53 samples (many of which were included in Kraiger and Ford (1985)) that contained both objective and subjective measures of job performance in their meta-analysis. Doing so allowed a comparison between race effects in the objective measures of job performance to race effects in the subjective measures of job performance. The logic of their study is that interpretations of race effects, like those found in Kraiger and Ford (1985), are not readily interpretable without some “ultimate” measure of job performance. In other words, comparing the relative size of race effects on objective measures of job performance to the size of race effects on subjective measures allows for easier interpretation of bias (Ford et al, 1986), assuming that the objective measures of job performance are unbiased themselves. Ford et al’s (1986) results indicated that overall, i.e. across objective and subjective measures of job performance, White ratees scored higher than Black ratees on measures of job performance. However, the magnitude of this effect was larger for the
subjective ratings of job performance. Ford et al (1986) concluded that although race effects are present in subjective measures of job performance, these race effects may be influenced by “true” differences in performance.

Martocchio and Whitener (1992) used 10 independent samples excluding studies using GATB data, to further explicate the issue of race effects in performance ratings, especially in relation to the Kraiger and Ford (1985) and Ford et al (1986) studies. With the understanding that these two previous meta-analyses concluded race effects present in supervisory ratings could be due to actual performance differences, Martocchio and Whitener (1992) suggested that the logic of Kraiger and Ford (1985) and Ford et al (1986) was potentially flawed. For instance, suggesting that race effects in supervisory ratings of job performance mirror race effects in objective measures of job performance makes the assumption that subjective and objective indices are measuring the same construct (Martocchio & Whitener, 1992). Therefore, Martocchio and Whitener (1992) addressed both the issue of rating bias, and the issue of convergent validity of objective and subjective ratings. Their results indicated that the validities of subjective and objective ratings, between cognitive ability, are highly similar (Martocchio & Whitener, 1992), suggesting convergent validity of objective and subjective measures of performance. Therefore, if subjective ratings of job performance are unbiased, ratee race effects for subjective ratings should be similar to the ratee race effects for objective measures of job performance. However, the differences between White ratees and Black ratees on subjective measures of job performance were much larger than their differences on objective measures, 0.284 for subjective and -0.009 for objective (Martocchio & Whitener, 1992). Martocchio and Whitener (1992) interpret this as evidence contrary to the conclusions of Kraiger and Ford (1985) and Ford et al (1986). Since race effects on subjective and objective measures of job performance were so
different and the race effects on subjective ratings were quite large, these results present evidence for rating bias in subjective measures of job performance.

Ford et. al. (1986) and Martocchio and Whitener (1992) both used objective measures as the metric by which to separate true ratee race performance differences from rating bias. Ford et. al. (1986) report smaller ratee race effects for objective ratings than subjective ratings and Martocchio and Whitener (1992) report no ratee race effect in their sample of objective measures. However, the common finding in the literature is the exact opposite, especially where job performance is saturated with task performance (e.g. Roth, Huffcutt & Bobko, 2003). Examining ratee race effects, Roth, et al (2003) report that comparisons between objective and subjective measures on various categories of job performance, which include quantity measures, quality measures, job knowledge and absenteeism, produced larger race effects for objective measures. The category of job knowledge produced a ratee race effect size of 0.55 for objective measures (i.e., tests of job knowledge) and 0.15 for subjective measures (i.e., ratings of job knowledge). Even absenteeism produced a ratee race effect size of 0.23 for objective measures and 0.13 for subjective measures. Due to small sample sizes, previous meta-analyses, like Ford et al (1986) and Martocchio and Whitener (1992), were forced to combine performance measures into “somewhat heterogeneous categories of performance (p. 694, Roth et al, 2003).” More homogeneous categories of job performance provide a more “accurate” understanding of any racial differences (Roth et al, 2003). Though this may provide insight as to why Ford et al (1986) and Martocchio and Whitener’s (1992) meta-analyses produced results counter to the literature, understanding the criterion of job performance provides further insight as to why one would expect race effects to be larger for objective versus subjective measures of job performance.
Using the same data from Pulakos et al (1989) (i.e. Project A data), Oppler et al (1992) examined the correlations between performance ratings and non-rating measures of job performance across rater race, ratee race, and rater race by ratee race groups. Examining these correlations, Oppler et al (1992) investigated the differences between those associated with the ratings provided by the White raters and those provided by the Black raters, for both White and Black ratees. The main goal of Oppler et al’s (1992) study, using the differential constructs approach, is to examine the degree to which performance ratings, associated with different rater race - ratee race combinations, correlate in the same manner to non-rating measures of performance. In other words, does rater or ratee race moderate the correlations between the performance ratings and non-rating measures of job performance? None of the analyses support the potential moderating effect of rater or ratee race on the correlations between supervisory ratings and non-rating measures of job performance (Oppler et al 1992). (Peer ratings did provide minimal evidence where this moderating effect was present.) However, Oppler et al’s (1992) results do indicate that “the degree of agreement between Black and White raters…was only slightly diminished after the removal of rating variance associated with the non-rating measures (p.217)”, indicating that the non-rating measures of job performance did not capture this systematic variation present in the supervisory ratings. Therefore, systematic differences still existed in the supervisory ratings of ratees after the variance related to the non-rating measures was removed. Oppler et al (1992) conclude that it is beyond the scope of their study to determine if this remaining systematic variance is due to other performance factors not included in the study or if it is due to other factors not related to performance, such as bias.

Summary. Under the assumption of no bias in performance ratings, objective measures of job performance and performance ratings should not be differentially affected by rater race,
ratee race or the interaction between rater race and ratee race; however, much of the research suggests they are differentially affected. More specifically, race effects associated with the use of non-rating or objective measures of job performance, where bias is expected to be minimal, and the race effects associated with the use of performance ratings differ in magnitude. Ford et al (1986) conclude that the differential effect sizes for the performance ratings and for the objective measures of performance do not indicate the existence of bias. Instead, the larger race effect associated with performance ratings are based on “true” performance differences evident in the objective measures of performance. Martocchio and Whitener (1992), despite concluding convergent validity between performance ratings and objective measures of job performance, interpreted the larger race effect associated with the performance ratings as bias. Unlike Ford et al (1986), Martocchio and Whitener (1992) conclude that the race effect associated with performance ratings was larger than expected based on the race effect associated with objective measures of performance, and must be due to bias in the performance ratings. The conclusions of both Ford et al (1986) and Martocchio and Whitener (1992), however, are suspect due to their findings of smaller ratee race effects for objective measures of job performance than ratee race effects for subjective measures of job performance, as mentioned previously. Oppler et al (1992) also presents evidence of convergent validity between performance ratings and objective ratings of job performance concluding the correlations between the performance ratings and objective ratings of job performance were not significantly affected by ratee race, rater race, or their interaction. Despite this, Oppler et al (1992) also present evidence indicating the existence of systematic differences associated with performance ratings beyond those accounted for by non-rating measures of performance.
Unlike the total association approach, the differential constructs approach is better suited to examine rating bias due to its ability to control true performance differences through the examination of convergent validity. However, the differential constructs approach rests on the assumption that performance ratings and objective measures of job performance are equivalent criteria of job performance. This assumption is not tenable. Research has provided evidence that the construct space of job performance is differentiated and that performance ratings and objective measures of job performance relate to different aspects of job performance (i.e., contextual and task performance) (e.g. Borman & Motowidlo, 1997; Saad & Sackett, 2002). More specifically, performance ratings are equally weighted task and contextual performance measures (Borman, White & Dorsey, 1995; Kiker & Motowidlo, 1999; Van Scotter & Motowidlo, 1996) whereas objective measures are primarily used as task performance measures (Schmidt, 1988). Though performance ratings and objective measures of job performance may correlate, they do not address the same aspects of job performance and, thus, are not equivalent criteria (Landy, 1989). This makes race effects difficult to interpret.

Direct Effects Approach

The direct effects approach attempts to “isolate the effects of subgroup membership that are not mediated by true performance differences between members of different subgroups (p. 203, Oppler et al, 1992).” Researchers attempt to statistically control for true performance differences between the subgroups and then examine any race effects. In essence, the goal of the direct effects approach is to control for true performance differences through the use of a predictor. Before conclusions are made using the direct effects approach, three assumptions must be met (Oppler et al, 1992). 1) The predictor or covariate used to account for the relationship between ratee race and the performance ratings (i.e. that represents true performance
differences) must represent a valid aspect of job performance. Therefore, any change to the relationship between the performance ratings and ratee race is assumed to be a function of true performance differences. 2) Any subgroup differences present on the predictor or covariate used must be attributable to true performance differences. In other words, the predictor or covariate used must be unbiased. 3) Even when the first two assumptions are met, the accuracy of interpreting any race effects depend on the assumption that no other factors related to true performance differences and ratee race exist. Failing to include all of the relevant predictors or covariates could result in the over or underestimation of race effects.

**Key direct effects study.** Oppler et al (1992) examined the race effects reported by Pulakos et al (1989) to determine if the relationship between performance ratings and ratee race was mediated by non-rating measures of performance. Therefore, Oppler et al (1992) conducted semi-partial correlations between performance ratings and ratee race, controlling for the variance associated with the non-rating measures of performance (i.e. four non-rating administrative measures and scores from training achievement tests). Oppler et al (1992) concluded that, generally, the zero-order correlations between the performance ratings and ratee race overestimated (underestimated) the relationship between performance ratings and ratee race, as provided by White (Black) raters. In other words, the correlations between performance ratings and ratee race became less positive (negative) for the White (Black) raters when comparing the semi-partial correlations to the zero-order correlations. Despite these changes in correlations, the rater race by ratee race interaction was not impacted (Oppler et al, 1992) and the initial conclusion of no bias in the performance ratings (Pulakos et al, 1989) remained unchanged. As noted previously, Pulakos (1989) concluded this significant rater race X ratee race interaction
was not indicative of rating bias due to the “trivial” amount of variance it accounts for in the performance rating scores.

*Differential prediction and the direct effects approach.* Differential prediction, which is a regression based analysis, is commonly employed to examine the predictive bias of tests. Predictive bias refers to the “consistent nonzero errors of prediction…made for members of the subgroup (p. 115, Cleary, 1968).” Assuming an unbiased criterion, differential prediction is used to determine if the prediction of some criterion is the same across groups (Cleary, 1968). If group prediction differs across sub-groups, predictor bias is concluded. In practice, testing predictive bias commonly uses performance ratings as the criteria (Landy, 1989); however, if performance ratings are biased, then interpretations of predictive bias are flawed. To test the bias in the criterion, i.e. performance ratings, it must be assumed that the predictor is free from bias and any subgroup differences are representative of true performance differences (i.e. assumptions 1 and 2 from Oppler et al, 1992). Differential prediction analyses can then be used to assess the existence of rating bias in performance ratings. Differential prediction is an alternative direct effects analysis to the semi-partial correlation approach used by Oppler et al (1992). Differential prediction can be used to examine performance ratings across ratees, as a function of raters (i.e., rater race X ratee race interaction), using the predictor to control for true performance differences, with the added advantage that the effects are expressed in the metric of the performance ratings. That is, instead of interpreting rating bias based on metric insensitive semi-partial correlations, differential prediction allows for the interpretation of rating bias based on adjusted mean criterion ratings.

When using differential prediction analyses to examine bias in performance ratings, it is assumed that the predictor is not biased. It is interesting that there is more evidence of racial
fairness for aptitude test scores than the racial fairness of performance ratings. There are several strategies for demonstrating the fairness of aptitude tests other than differential prediction studies (cf. Jensen, 1980; Herrnstein & Murray, 1994). Research consistently reports subgroup differences in overall cognitive ability test scores. White and Black test takers tend to exhibit a one standard deviation (SD) difference, where Whites receive higher scores (Hartigan & Wigdor, 1989; Herrnstein & Murray, 1994). Researchers have long been interested in the source of group differences resulting from these tests. Jensen (1980) and Herrnstein and Murray (1994) both posit that most of the findings in the literature either fail to support or contradict the expectations that cognitive ability tests are culturally biased against Black test takers. For instance, despite a smaller proportion of Black test takers answering “easy” and “hard” questions correctly (resulting in subgroup mean differences in overall test scores where Blacks score lower than Whites) the order of difficulty is virtually the same across subgroups (Herrnstein & Murray, 1994). The largest subgroup differences existed on items judged to be the least culturally-loaded (Jensen, 1980). Jensen (1980) argued that this is due to the g-loading associated with such items. Those items that are the least culturally-loaded are more highly loaded on the g factor (i.e. mental ability such as reasoning). Therefore, “the white-black difference [in test scores] is mainly a difference in g rather than in group factors that are specific to certain types of items (p. 585, Jensen, 1980).”

Key Differential Prediction Studies. Rotundo and Sackett’s (1999) study attempted to address a concern that the common finding of no differential prediction is due to a biased supervisory ratings. Rotundo and Sackett (1999) used data from the GATB validation research conducted between 1972 and 1987, which is the same data included in Sackett and DuBois (1991), Martocchio and Whitener (1992), and Stauffer and Buckley (2005). Therefore, Rotundo
and Sackett (1999) conducted differential prediction analyses to compare ability-performance relationships under two conditions: when all employees (i.e., both Black and White ratees) were rated by a supervisor from the majority group (i.e., White rater) and when all employees were rated by a supervisor from their same race. If the analyses of performance ratings for employees who were rated by a supervisor from their same race do not affect overall conclusions regarding predictive bias, as compared to the performance ratings for employees who were rated by a supervisor from the majority group, then the concern of a biased criterion becomes “less tenable” (Rotundo & Sackett, 1999). This rests on Rotundo and Sackett’s (1999) conclusions that ratings from supervisors of the same race provide less fodder for claims of bias. Conducting both between-subjects and within-subjects comparisons, matched rater-ratee race analyses were compared to majority group rater analyses. Rotundo and Sackett (1999) concluded that neither majority group rater analyses, i.e. between-subjects or within-subjects, indicated the under prediction of the minority group. When compared to the matched rater-ratee race analyses, where no evidence for predictive bias existed, results suggested that conclusions regarding predictive bias did not change as a function of rater race. Rotundo and Sackett (1999) concluded no evidence for rating bias in performance ratings.

However, Rotundo and Sackett (1999) failed to conclude evidence for rating bias in performance ratings when their results indicated its existence (Hauenstein and Tison, 2007). The between-subjects majority group rater analyses indicated intercept bias where Black (White) ratee performance was overpredicted (underpredicted) by the common regression line (Rotundo & Sackett, 1999). When the between-subjects matched rater-ratee race analyses were then examined, this intercept bias disappeared. Rotundo and Sackett (1999) were only concerned with determining if rater bias was an explanation for the absence of differential prediction, not if
rater bias was an explanation for the disappearance of differential prediction. Additionally, Rotundo and Sackett (1999) also failed to model the full rater race X ratee race comparisons, examining only matched rater race – ratee race and majority rater – ratee groupings. It is with examination of the full model where differential prediction is the most useful in the detection of rating bias in performance ratings. Modeling the full interaction provides a metric by which to examine sub-group means and make decisions regarding the practical importance of their differences.

Waldman and Avolio (1991) attempted to address the caveat that evidence of racial bias in supervisory ratings exists mainly in laboratory settings (Dipboye, 1985) by providing evidence of race effects in a field setting. Using data from the United States Employment Service of the Department of Labor (i.e. GATB data), Waldman and Avolio (1991) conducted regression analyses on supervisory ratings of 21,547 ratees (14,403 White rater – White ratee, 5,699 White rater- Black ratee, 529 Black rater – White ratee, 916 Black rater – Black ratee). Their results indicated that neither the ratee race nor the rater race contributed significantly to the prediction of performance ratings. Additionally, Waldman and Avolio (1991) presented no evidence of the rater race X ratee race interaction effect and conclude that “little or no bias in supervisory ratings” exists (pg. 901). However, Waldman and Avolio (1991) stress caution in interpreting these results due to the small sample sizes associated with the Black raters (i.e. Black rater – White ratee and Black rater – Black ratee conditions). Small sample sizes reduce the power to detect significant effects (Pedhauzer & Schmelkin, 1991).

Hauenstein and Tison (2007) presented an additional benefit to the use of differential prediction: its assistance in determining the source of rating bias. As previously discussed, research has provided evidence that the construct space of job performance is differentiated (i.e.,
contextual and task performance) and that performance ratings and objective measures of job performance relate to different aspects of job performance (e.g., Borman & Motowidlo, 1997; Saad & Sackett, 2002). Subjective ratings of job performance contain equal weightings of task and contextual measures (Borman, White & Dorsey, 1995; Kiker & Motowidlo, 1999) and objective measures are primarily used as task performance measures (Schmidt, 1988). Research has also provided evidence that ratee race effects are, on average, 0.40 SD for task measures of job performance (e.g., Roth et al, 2003; McKay & McDaniel, 2006) and 0.20 SD for contextual measures of job performance (e.g., Hauenstein et al, 2002; McKay & McDaniel, 2006). If subjective measures of job performance are equally weighted task and contextual measures, where expected ratee race effects are 0.40 SD for task measures and 0.20 SD for contextual measures, then supervisory ratings should display an average ratee race effect of 0.30 SD (Hauenstein & Tison, 2007).

Hauenstein and Tison (2007) used this expected ratee race effect of 0.30 SD and re-analyzed research findings (i.e., Hartigan & Wigdor, 1989; Roth et al, 2003; Stauffer & Buckley, 2005) to determine when ratee race effects did not match the expected 0.30 SD difference. The logic was, if White and / or Black raters produced a ratee race effect larger (smaller) than the expected 0.30 SD, then they were rating too severely (leniently). In Stauffer and Buckley (2005), the ratee race effects for White raters mirrored the 0.30 SD expected effect size but the ratee race effects for Black raters did not (i.e. 0.044 SD). Given that the White and Black raters’ ratings of White ratees were similar and their ratings of Black ratees were different, Hauenstein and Tison (2007) concluded that this finding indicates Black raters were rating Black ratees too leniently. They stress, however, that this interpretation of bias in Black raters’ performance ratings is not always the case. For instance, in examining the ratee race effect sizes reported in
Hartigan and Wigdor (1989) and Roth et al (2003), which report on the same GATB data, White raters were determined to be rating Black ratees too severely (Hauenstein & Tison, 2007). It is important to note that this also suggests that the GATB data examined indicates two forms of bias in the performance ratings: Black raters rating Black ratees too leniently (i.e. Stauffer & Buckley, 2005) and White raters rating Black ratees too severely (i.e., Hartigan & Wigdor, 1989, and Roth et al, 2003).

Summary. As an example of the direct effects approach, the use of differential prediction can be quite useful in the examination of rating bias. Re-examining Roth et al (2003), Hauenstein and Tison (2007) pin-pointed that when correctly utilized differential prediction can be a useful tool in detecting instances of rating bias. Furthermore, examination of the full model can also aid in assertions regarding the source of rating bias (Hauenstein & Tison, 2007). Race effects are also expressed in the metric of the performance ratings. This allows better assertions regarding the practical significance of the race effects. To date, however, differential prediction models do not commonly examine the full rater race by ratee race interaction. One potential reason for this is that modeling the full rater race by ratee race interaction may lack sufficient power due to the small number of Black ratees available (Aguinis, 1995; Aguinis & Stone-Romero, 1997) and, more importantly, the small number of Black raters rating Black ratees. Waldman and Avolio (1991) exemplify the downfall of small sample sizes: little to no significant effects.

Overview

Research using the total association and differential constructs approaches provide limited ability to examine the existence of rating bias in performance ratings. In the total
association approach, true performance is confounded with ratee race effects. Despite controlling for true performance effects by examining the convergent validity of performance ratings and objective measures of job performance, the differential constructs approach makes an unfounded assumption. It assumes that the performance ratings and objective measures of job performance are equivalent criteria, when in fact they are not. The direct effects approach, and more specifically differential prediction analysis, is more useful than the total association and differential constructs approaches. By using a predictor to control for true performance differences, performance ratings can be examined for evidence of rating bias. Issues of practical significance are also examined given that race effects are expressed on the metric of the performance ratings. The usefulness of differential prediction depends, however, on modeling the full rater race X ratee race interaction. Issues of power, due to small sample sizes, is one reason that modeling this full interaction is not currently implemented in the literature.

Purpose

This study uses monte carlo simulations to model the ratee race X rater race interaction and examine the conditions where differential prediction has sufficient power to be a useful tool in the detection of rating bias. The variables manipulated to determine optimal conditions for detecting rating bias include: the total sample size of ratees, the proportion of White and Black raters, the proportion of White and Black ratees, the validity of the predictor scores, and the magnitude of rating bias present in the performance rating scores.
Methods

Simulation Program

A SAS program was constructed to create population data of the ratees, as a function of the manipulated parameters (see below for a description of the manipulated parameters), that were used in determining the power of detecting a significant rater race X ratee race interaction. To do this, the SAS program created a population distribution of 10,000 cases for two vectors, a standardized predictor score and a standardized criterion score. In each iteration, each vector was normally distributed with a mean of zero and the two vectors were correlated at a specified validity. Next, 50% percent of the cases were designated as Black ratees and 50% were designated as White ratees. Within each ratee race population, 50% (i.e., N = 2,500) were designated as rated by Black raters and 50% were designated as rated by White raters. For White ratees, +1 was added to the standardized predictor score to reflect a one standard deviation difference between White and Black ratees on aptitude scores. For rating bias, the pattern of Black ratee bias reported by Stauffer and Buckley (2005) was used\(^1\). This indicates that the magnitude of bias examined was attributable to the Black ratee - Black rater group. (Despite attributing bias to the Black ratee – Black rater group, this does not indicate the source of the rating bias.) To capture this pattern of rating bias, the criterion scores of the 7,500 cases that do not represent Black ratees rated by White raters were adjusted by adding a standard deviation constant (e.g., 0.30) that reflected the presence of rating bias.

Hartigan and Wigdor (1989) reported only two incidents of slope bias out of 72 (3%) validation studies (i.e. GATB validation studies). Moreover, traditional differential prediction research detects very few instances of slope bias. Therefore, to aid in clarity, it was assumed that
slope bias was not present for these analyses. Under this assumption, the effects of rating bias can be demonstrated as the presence / absence of intercept bias using differential predication analyses.

*Manipulated Parameters*

Manipulating total sample size is a common way to adjust power (e.g., Aquinis, 1995). However, the inclusion of rater race proportion and ratee race proportion complicates the issue. The small number of Black ratees available (Aguinis, 1995; Aguinis & Stone-Romero, 1997) and the small number of Black raters rating Black ratees are a major concern regarding the power issues of differential prediction analyses. Conditions of small sample sizes, especially those in the Black rater – Black ratee group, result in little to no significant effects (Waldman & Avolio, 1991). Therefore, the boundary conditions have been designed (i.e. the total sample size, the proportion of raters and the proportion of ratees) to produce, at a minimum, a sample size of fifteen in the Black rater – Black ratee condition.

With the assumption of no slope bias, both magnitude of rating bias and validity of predictor scores directly alter the y-intercepts of rater race – ratee race subgroups (see Hauenstein & Tison, 2007), thus impacting power. Manipulations of these parameters were included to assess how these alterations impact power levels and to provide information regarding their generalizability. Understanding generalizability helps to address such questions as: when the magnitude of rating bias is small (i.e., \( d = 0.15 \)), under what conditions is there adequate power to detect a significant rater race X ratee race interaction?
All parameters had three manipulated values in this study. Doing so allowed for linear inspection of the relationship between these parameters and the proportion of significant rater race by ratee race interactions detected, i.e., power levels.

Validity coefficient. The simulation generated the population distribution for each ratee group by using the correlation between the predictor scores and the generated performance rating scores, i.e., the validity coefficient. Three different conditions were analyzed: low validity (0.10), moderate validity (0.30), and high validity (0.50).

Magnitude of rating bias. Rater race – ratee race subgroup means on the generated performance rating scores were adjusted by the magnitude of rating bias. The magnitude of rating bias was examined as three effect sizes: weak (0.15), moderate (0.30), and strong (0.45).

Total sample size. This indicates the total sample size, i.e. total number of White and Black ratees, examined for the existence of rating bias. Three different conditions were analyzed: small (500), medium (1,000), and large (2,000).

Rater race proportion. The proportion of White and Black raters used to sample ratee scores were 70 / 30, 80 / 20, and 90 / 10, where the first number indicates the proportion of White raters and the second number indicates the proportion of Black raters.

Ratee race proportion. The proportion of White and Black ratee scores sampled were 50 / 50, 60 / 40, 70 / 30, where the first number indicates the proportion of White ratees and the second number indicates the proportion of Black ratees.
Analyses

For each combination of parameters in the SAS simulation, the full ratee race X rater race differential prediction analysis was replicated 1,000 times. The dependent variable of interest was the percentage of time the ratee race X rater race interaction was significant, using 0.05 as the Type I error rate, across these 1,000 iterations. The combinations of parameters produced 243 (i.e., $3^5$) simulation runs of 1,000 differential prediction analyses. This, in turn, produced 243 data points, i.e., the percentage of times the three-interaction was significant in each run of 1,000 analyses. The data points were examined to determine which combination of parameters produced acceptable levels of power, both minimal and traditional (i.e., 0.60 and 0.80).

Using SPSS, a components analysis of variance was then conducted on all 243 data points to determine which parameters (and interactions between parameters) drive the variation in the power levels. This analysis partitioned variance into each source of variation (i.e., the main effect of each manipulated parameter and the interactions among the parameters, up to the 4-way interactions).

Results

The purpose of this study was to examine the effects of total sample size, magnitude of rating bias, validity of predictor scores, rater race proportions and ratee race proportions on the power to detect a significant rater race X ratee race interaction in differential prediction analyses. 

Power Analysis

For clarity, the results of the power analyses are broken down by the magnitude of the rating bias.
Small rating bias. Small rating bias is represented as $d = 0.15$. For small rating bias, the maximum power is 0.32. This occurs when sample size is 2,000, validity of predictor scores is 0.50, rater race proportion is 70 / 30, and ratee race proportion is 50 / 50. It is clear that when rating bias is small, differential prediction is not a practical method of detecting a rater race $X$ ratee race interaction.

Moderate Rating Bias. Moderate rating bias is represented as $d = 0.30$. When sample size is 500, the maximum power is 0.30 for the most favorable combination of parameters. Power begins to approach 0.60 when sample size is 1,000. To show this more clearly, Figure 1, Figure 2, and Figure 3 present the power results for sample size equal to 1,000. In Figure 1, validity equals 0.10, and the bars in the histogram represent the nine combinations of the rater race proportions and ratee race proportions. Validity equals 0.30 in Figure 2, and validity equals 0.50 in Figure 3. Figure 3 shows that the maximum power achieved for sample size equal to 1,000 (i.e., power equals 0.56) occurs when validity equals 0.5, rater race proportion is 70 / 30, and ratee race proportion is 60 / 40.

With moderate rating bias, power reaches minimal acceptable levels (i.e., 0.60) when sample size equals 2,000. Figure 4, Figure 5, and Figure 6 present the power results for a sample size equal to 2,000, where validity equals 0.10 in Figure 4, 0.30 in Figure 5 and 0.50 in Figure 6. Figure 4 shows that power reaches minimally acceptable levels when validity equals 0.10 and the rater race proportion is 70 / 30. Also, the minimum criterion is met when validity equals 0.10, rater race proportion is 80 / 20, and ratee race proportion is 60 / 40. Figure 5 shows that when validity increases to 0.30, the change from when validity equals 0.10 is that now all 80 / 20 rater race proportions produce minimum acceptable power. Finally, Figure 6 shows that when validity
increases to 0.50, the 70 / 30 rater race proportions reach the traditional acceptable power level (i.e., 0.80).

*Large Rating Bias.* Large rating bias is represented as $d = 0.45$. When sample size is equal to 500, the most favorable combination of parameters still do not produce the minimal acceptable level of power. Figure 7, Figure 8, and Figure 9 present the power results for a sample size of 500 with a validity of 0.10 in Figure 7, a validity of 0.30 in Figure 8, and a validity of 0.50 in Figure 9. Figure 9 shows that the maximum power achieved for a sample size of 500 is 0.58. This occurs when validity equals 0.50, rater race proportion is 70 / 30 and ratee race proportion is 50 / 50.

Unlike the conditions with moderate rating bias, conditions with large rating bias reaches minimal acceptable levels, and even traditional acceptable levels, of power when sample size equals 1,000. This is shown more clearly in Figure 10, Figure 11, and Figure 12, where the power results are presented for sample size equal to 1,000. In Figure 10, validity equals 0.10. Validity equals 0.30 in Figure 11, and validity equals 0.50 in Figure 12. Figure 10 shows that power reaches minimally acceptable levels when validity equals 0.10 and the rater race proportion is 80 / 20, as well as when validity equals 0.10 and the rater race proportion is 70 / 30. The results are similar in Figure 11, except now, when validity increases to 0.30, power reaches traditional acceptable levels when rater race proportion is 70 / 30 and ratee race proportion is 60 / 40 or 50 / 50. Figure 12 shows that when validity increases to 0.50, all 70 / 30 rater race proportions reach the traditional acceptable level of power.

When sample size equals 2,000, power reaches either the minimal acceptable level or the traditional acceptable level in all conditions, with large rating bias. Figure 13, Figure 14, and Figure 15 present the power results for a sample size equal to 2,000, where validity equals 0.10
in Figure 13, 0.30 in Figure 14 and 0.50 in Figure 15. Figure 13 shows that power reaches minimally acceptable levels when validity equals 0.10 and the rater race proportion is 90 / 10. Additionally, the traditional criterion is met when validity equals 0.10 and rater race proportion is 80 / 20 or 70 / 30. Figure 14 shows that when validity increases to 0.30, the results mirror those in Figure 13. That is, the conclusions of which conditions reach the minimal and traditional acceptable levels of power remain the same across Figure 13 and Figure 14. In Figure 15, when validity increases to 0.50, all conditions reach the traditional acceptable power level.

Summary. No combinations of parameters reach acceptable levels of power when the magnitude of rating bias is small. Even when the magnitude of rating bias increases to 0.30, combinations of parameters do not reach acceptable power levels until sample size is 2,000. Combinations of parameters with a large magnitude of rating bias produce the most acceptable power levels. In these conditions, acceptable power can be obtained when sample size is just 1,000.

Components of Variance Analyses

In addition to knowing what combination of parameters reach acceptable levels of power, it is important to determine which parameters (or interactions between parameters) drive the variation in the power levels.

Main effects. Table 1 represents the main effects for the full components analysis of variance. Magnitude of rating bias accounts for the most variance in power levels (i.e., 53.21%) followed by sample size (i.e., 24.27%) and rater race proportions (i.e., 8.40%). Neither validity of predictor scores nor ratee race proportion accounts for more than 1% of the variance in power levels. Figures 1 thru 15 allow for a visual examination of how these parameters impact power
levels. Given that the majority of variance in power levels is accounted for by magnitude of rating bias, sample size, and rater race proportion, these effects are explored below.

The effect of the magnitude of rating bias on power levels is exemplified by the comparison of Figure 3 to Figure 12. These two figures differ only in terms of magnitude of bias (moderate versus large). Collapsing across all ratee race and rater race proportions, the average difference in power is 0.40, between moderate and large bias. Collapsing over all other parameters, the average power for small, medium, and large bias was, 0.14, 0.39, and 0.64, respectively.

In terms of the effect of sample size, Figure 3 and Figure 6 represent combinations of parameters where only sample size differs (i.e., 1,000 versus 2,000). There is an average increase in power of 26% when sample size increases from 1,000 to 2,000, collapsing over the ratee race and rater race proportions. Collapsing over all other parameters, the average power for 500, 1,000, and 2,000 sample sizes were 0.22, 0.38, and 0.57, respectively.

Finally, for the rater race proportion main effect, examination of any figure exemplifies the effect on power. Using Figure 3 again, the average increase in power is 18% from 90 / 10 to 80 / 20, and 9% from 80 / 20 to 70 / 30. Collapsing across all other parameters, the average power for 90 / 10 80 / 20, and 70 / 30 were 0.28, 0.41, and 0.48, respectively.

*Interactions.* The results from the full components of variance analysis indicated three interactions that accounted for a small portion of the variance in power levels: total sample size X magnitude of rating bias (9.33%), magnitude of rating bias X rater race proportion (2.80%), and total sample size X magnitude of rating bias X rater race proportion (0.93%). Given that the two two-way interactions are included in the three-way interaction, the three-way interaction was of most interest. Total sample size as a main effect, however, accounts for a sizable amount of
variance in power levels (i.e., 24.27%). Therefore, to better understand the three-way interaction, the components of variance analysis was re-run for each total sample size (i.e., 500, 1,000, 2,000). The magnitude of rating bias X rater race proportion interaction was then examined across these individual runs. This produced a better understanding of how these three parameters interact to effect power levels.

For the small sample size (see Table 2), the magnitude of rating bias X rater race proportion interaction accounted for 13.73% of the variance in power levels. For the moderate sample size (see Table 3), the rating bias X rater race proportions captured 6.01% of the variance. Finally, for the large sample size (see Table 4), the rating bias X rater race proportions captured 2.29% of the variance.

Table 5 represents the power levels associated with the magnitude of rating bias X rater race proportion interaction, collapsed over all other parameters (i.e., total sample size, validity of predictor scores and ratee race proportions). For a small magnitude of rating bias (i.e., 0.15), changing rater race proportion from 90 / 10 to 80 / 20 increases power levels by 16% (i.e., 0.26 – 0.10) and changing rater race proportion from 80 / 20 to 70 / 30 increases power levels by 22% (i.e., 0.48 – 0.26). In contrast, when the magnitude of rating bias is moderate and large, the same increases in the power levels were similar as the rater race proportion increased. For moderate and large rating bias, the increases in power were all between 26% and 28%.

Table 6 represents the power results for the total sample size X magnitude of rating bias X rater race proportion interaction. (The power levels presented in Table 6 are collapsed over validity of predictor scores and ratee race proportions.) For sample sizes of 500 and 1,000, the pattern is that the increases in power as a function of the rater race proportion effect are greatest when rating bias is strong. The increases in power are greatest when rating bias is moderate
when the sample size is 2,000. This change in the pattern is caused by a ceiling effect when the largest sample size is combined with the largest rating bias effect. For the sample size of 2,000, when the magnitude of rating bias is 0.30, changing rater race proportion from 90 / 10 increases power from 0.42 to 0.65 (i.e., 23%) and changing rater race proportion from 80 / 20 to 70 / 30 increases power from 0.65 to 0.76 (i.e., 11%). However, the same changes to rater race proportion, when rating bias is 0.45, only increases power 19% (i.e., 0.93 – 0.74) and 5% (0.98 – 0.93).

Discussion

Previous research examining rating bias in performance ratings has employed three common techniques: the total association approach, the differential constructs approach and the direct effects approach. Upon reviewing the literature, the direct effects approach, and more specifically differential prediction analysis, is determined to be more useful in the detection of rating bias. Recall that 1) true performance is confounded with ratee race effects in the total association approach and 2) the differential constructs approach, despite attempting to control for true performance effects, rests on the unfounded assumption that performance ratings and objective measures of job performance are equivalent criteria. With differential prediction analyses, a predictor is used to control for true performance differences allowing performance ratings to be examined for evidence of rating bias. Race effects are also expressed in the metric of the performance ratings, allowing for better assertions regarding the practical significance of their differences. However, the usefulness of differential prediction analyses depend on modeling the full rater race X ratee race interaction and having adequate power to detect it.

The purpose of this study was to explore how sample size, magnitude of rating bias, validity of predictor score, rater race proportions and ratee race proportions impact the power to
detect the rater race X ratee race interaction, using differential prediction analyses. Moreover, this study sought to determine under which conditions these parameters and / or combinations of these parameters produce adequate power to detect the rater race X ratee race interaction. The differential prediction analyses provided the power levels for the different combination of parameters. These results were presented to provide an understanding of when adequate power levels are achieved; however, these results do not provide insight as to which parameters are driving the changes in power levels. Using the components of variance analyses, the proportion of variance in power levels accounted for by parameters (and interactions between parameters) was identified. This discussion will examine both the main effects and the interaction effects represented in the data. Combining the results from the differential prediction and components of variance analyses, this study provides both researchers and practitioners general guidance on how these parameters impact power levels and when the use of differential prediction analyses provide adequate power in the detection of rating bias in performance ratings.

Table 7 is included to simplify the discussion of the differential prediction results and is representative of the power levels provided in Figures 1 thru 15. Here, the power levels for all sample size, magnitude of rating bias, validity of predictor scores, and rater race proportion combinations, collapsed over ratee race proportions. (An explanation as to why power levels were collapsed over ratee race proportion is included in the following section.) The combinations of parameters that produce adequate power, both minimal and traditional power levels, are represented in bolded font.

Main Effects

Ratee race proportion is not included in this table, or in this discussion, for two reasons: 1) it does not account for any meaningful amount of variance in power levels and 2) the slight
fluctuations it causes in power levels requires additional research to fully understand. In the components of variance analysis, ratee race proportion only accounted for 0.03% of the variance in power levels. Moreover, it did not account for any more than 0.03% in any of its interactions with the other parameters. In short, no significant amount of variance was accounted for by ratee race proportion or any interaction with ratee race proportion. However, when examining the power levels from the differential prediction analyses, ratee race proportions did produce slight variations in recorded power levels. The general trend is that power levels increase as ratee race proportion changes from 70 / 30 to 60 / 40, but then drops as ratee race proportion changes from 60 / 40 to 50 / 50. Recall that the power to detect rating bias using differential prediction, in part, relies on the sample size in each of the four sub-groups (i.e., White rater – White ratee, White rater – Black ratee, Black rater – White ratee, and Black rater – Black ratee). For each total sample size (i.e., 500, 1000, and 2000), there is an optimal parsing of ratees into these groups. As ratee proportions change, deviations from this optimal parsing cause power to fluctuate. More research is needed to better understand how the designation of ratees into the four sub-groups effects power. However, for the purposes of this study, these fluctuations in power levels were not large and are not included in this discussion. The other parameters are collapsed over ratee race proportion and their effects on power levels are discussed below.

Magnitude of rating bias accounts for the largest amount of variance in power levels (i.e., 53.21%). Without considering any other parameters, this indicates that the magnitude of bias drives the variation in power levels. As discussed previously, this manifests itself in the following way: as the magnitude of rating bias increases, power levels increase. An important question, however, is at what magnitude of rating bias is there adequate power to detect it? In Table 7, when the $d = 0.15$ there are no instances where power reaches an acceptable level,
indicating that when the magnitude of rating bias is small, there is little chance of detecting it. However, power does begin to reach acceptable levels when \( d = 0.30 \). It is around this point that researchers and practitioners can begin to have confidence in differential prediction analyses as a tool to detect rating bias. Increasing to \( d = 0.45 \) provides even more instances where power begins to reach acceptable levels. Though researchers and practitioners may not know the magnitude of bias present in practice, they can obtain estimates of rating bias to use as a guide by examining race differences in performance rating scores. This provides an estimate of rating bias to aid in determining if differential prediction would be useful. In essence, these estimates (i.e., race differences in performance rating scores) are the amount of rating bias present if the validity of the predictor scores is equal to zero. Therefore, when obtained estimates of rating bias approach 0.30 and beyond, differential prediction would be useful and should be considered as a tool to use in detecting rating bias.

Similar to magnitude of rating bias, total sample size accounts for a large amount of variance in power levels (i.e., 24.27%). Again, without considering any other parameters, this indicates that the total sample size drives a large portion of the variation in power levels. Table 7 shows that as sample size increases, power levels also increase. Adequate power levels, however, are not obtained when sample sizes are small (i.e., 500). Instead, power levels begin to reach acceptable levels once sample sizes are moderate to large (i.e., 1000 to 2000) indicating that, in practice, differential prediction will be the most useful when sample sizes are moderate to large.

As seen in Table 7, as rater race proportion changes to include a larger proportion of Black raters, power levels increase. Given that rater race proportion accounts for 8.40% of variance in power levels (see Table 1), it is clear that this is a key parameter to consider as well.
Moreover, each level of rater race proportion reaches acceptable levels of power, though none of these appear when total sample size or magnitude of rating bias is small. It’s important to note, however, that there is a larger increase in power levels as rater race proportion changes from 90 / 10 to 80 / 20 (i.e., 13% on average, collapsed over all other parameters) than when it changes from 80 / 20 to 70 / 30 (i.e., 7% on average, collapsed over all other parameters). Therefore, increasing the proportion of Black raters to be 20% and higher provides more confidence in detecting the rater race X ratee race interaction in differential prediction analyses.

Although the validity of predictor scores accounted for less than 1% of the variance in power levels (i.e., 0.93%), a discussion of this parameter is still warranted. In Table 7, a distinct trend in power levels is shown; as validity of predictor scores increase, power levels also increase. Also, acceptable power levels are obtained in each of the three levels of predictor score validities (i.e., 0.10, 0.30, and 0.50). (Similar to that of rater race proportions, no acceptable power levels are obtained when total sample size or magnitude of rating bias is small.) Despite achieving acceptable levels of power in all three levels, more are achieved as validity of predictor scores increase. Therefore, as validity of predictor scores increase beyond 0.10, differential prediction analyses become a more useful tool in the detection of rating bias.

Interactions

Though examination of the parameters individually provides guidance as to when they are the most useful in achieving acceptable levels of power, the most informative aspect of this study is in determining how the parameter interactions effect power levels. The components of variance analysis on the complete set of differential prediction data (i.e., 243 data points), indicated three interactions: total sample size X magnitude of rating bias, magnitude of rating bias X rater race proportion, and total sample size X magnitude of rating bias X rater race
proportion. As previously stated, the three-way interaction was of most interest given that the other two-way interactions were included in it. However, given that total sample size accounts for large portion of variance in power levels (i.e., 24.27%), the components of variance analysis was re-run for each total sample size (i.e., 500, 1,000 and 2,000). The only interaction that accounted for a significant amount of variance in power levels, upon examining these analyses, was the magnitude of rating bias X rater race proportion interaction. Although this interaction accounts for only 0.93% of variance in power levels (see Table 1) for the total sample, it accounts for 13.73%, 6.01%, 2.29% when sample size is 500, 1000 and 2000, respectively.

Table 5 represents how the magnitude of rating bias X rater race proportion interaction effects power levels: increasing both magnitude of rating bias and the proportion of Black raters increases power levels beyond that of just increasing one of the parameters alone. This trend holds regardless of the total sample size (see Table 6). Table 6 represents how the total sample size X magnitude of rating bias X rater race proportion interaction effects power levels. Similar to the two-way interaction mentioned above, increasing the total sample size, the magnitude of rating bias, and the proportion of Black raters increases power levels beyond that of just increasing any one or two-parameter combination of these parameters. It is interesting, however, that as sample size increases, the proportion of variance in power levels accounted for by this three-way interaction decreases. Table 6 allows for a better examination of this.

In situations where total sample size is large and the magnitude of rating bias is large, the proportion of Black raters is not that important. These conditions produce acceptable levels of power regardless of the rater race proportion. However, if sample size or magnitude of rating bias decreases, higher proportions of Black raters are needed to produce acceptable levels of power. In practice, if estimates of rating bias are large and sample sizes are large, regardless of
the other parameter values, differential prediction analysis would be a useful tool to detect the
erater race X ratee race interaction. However, if total sample size (magnitude of rating bias) is
moderate, the magnitude of rating bias (total sample size) must be large in order to consider
differential prediction a useful tool. Moreover, in these conditions, the proportion of Black raters
should be approaching 20% and beyond.

Summary

In examining the individual parameters, their most useful, or optimal, values were
identified: sample size of around 1000, magnitude of bias around 0.30, and rater race proportion
of 80 / 20 and above. Table 5 and Table 6 provide more information regarding how total sample
size, magnitude of rating bias and rater race proportion interact to impact power levels.
Considering only these three parameters, as long as the proportion of Black raters is 20% or
above, total sample size (magnitude of rating bias) that is moderate and magnitude of rating bias
(total sample size) that is large will produce acceptable levels of power. In these cases,
differential prediction is considered a useful tool. Table 7 mirrors the results of Table 6 even
with the inclusion of validity of predictor scores, indicating that the parameters that have the
most impact on power levels are total sample size, magnitude of rating bias and rater race
proportion.

If either total sample size or magnitude of rating bias is small, no combination of
parameters produces acceptable levels of power. Therefore, differential prediction analyses
should not be used as the method to detect rating bias, as it does not have sufficient power.
Researchers and practitioners have adequate power to use differential prediction analyses when
both total sample size and magnitude of rating bias are large, regardless the value of the other
parameters. However, if either the total sample size or the magnitude of rating bias is moderate,
then rater race proportion becomes an important parameter. In these instances, the proportion of Black raters need to be around 20% or higher to have confidence in differential prediction analyses as a tool for detecting rating bias.

Conclusions

This study highlights the advantages of using differential prediction in examining the existence of rating bias in performance ratings. By using a predictor to control for true performance differences and providing race effects on the metric of performance ratings, differential prediction analyses overcome limitations of other methods (i.e., the total association approach and the differential constructs approach). Despite the advantages, however, differential prediction is not widely applied. This is due, in part, to issues related to power. (Recall that the usefulness of differential prediction rests with modeling the full rater race X ratee race interaction and small sample sizes impact the power to detect this interaction.) This study used monte carlo simulations to assess the power of differential prediction analyses in detecting small, moderate and large amounts of rating bias. In doing so, this study informs researchers and practitioners as to when differential prediction has adequate power to detect the rater race X ratee race interaction and, thus, becomes a useful tool in the examination of rating bias.
Footnote

1Kraiger and Ford (1985) also argued for a pattern of rater bias where each rater race favors ratees of their own race, i.e., same race bias. Hauenstein and Tison (2007) conducted analyses for this same race bias in addition to the Black ratee bias suggested by Stauffer and Buckley (2005). They found that, “relative to the Black ratee bias pattern, the same race bias pattern is more likely to lead to intercept bias where Black (White) ratee performance is underpredicted (overpredicted) by the common regression line (p. 25, Hauenstein & Tison, 2007).” I chose to use the pattern of bias suggested by Stauffer and Buckley (2005), given that the intercept bias represented by the same race bias is such a rare finding in the literature.
References


Table 1

*Proportion of Variance Accounted for in Power Levels by the Manipulated Parameters*

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude of Rating Bias</td>
<td>53.21%</td>
</tr>
<tr>
<td>Total Sample Size</td>
<td>24.27%</td>
</tr>
<tr>
<td>Rater Race Proportion</td>
<td>8.40%</td>
</tr>
<tr>
<td>Validity of Predictor Scores</td>
<td>0.93%</td>
</tr>
<tr>
<td>Ratee Race Proportion</td>
<td>0.03%</td>
</tr>
</tbody>
</table>
Table 2

*Proportion of Variance Accounted for in Power Levels by the Main Effects and Interaction Effects of the Manipulated Parameters for a Small Sample Size (n = 500)*

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude of Rating Bias (Rating Bias)</td>
<td>72.10%</td>
</tr>
<tr>
<td>Validity of Predictor Scores (Validity)</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rater Race Proportion</td>
<td>13.73%</td>
</tr>
<tr>
<td>Ratee Race Proportion</td>
<td>0.08%</td>
</tr>
<tr>
<td>Rating Bias X Validity</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias X Rater Race Proportion</td>
<td>13.73%</td>
</tr>
<tr>
<td>Rating Bias X Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Validity X Rater Race Proportion</td>
<td>0.15%</td>
</tr>
<tr>
<td>Validity X Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rater Race Proportion X Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias X Validity X Rater Race Proportion</td>
<td>0.13%</td>
</tr>
<tr>
<td>Rating Bias X Validity X Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias X Rater Race Proportion X Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Validity X Rater Race Proportion X Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias X Validity X Rater Race Proportion X Ratee Race Proportion</td>
<td>0.06%</td>
</tr>
</tbody>
</table>
Table 3

*Proportion of Variance Accounted for in Power Levels by the Main Effects and Interaction Effects of the Manipulated Parameters for a Medium Sample Size (n = 1000)*

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude of Rating Bias (Rating Bias)</td>
<td>76.91%</td>
</tr>
<tr>
<td>Validity of Predictor Scores (Validity)</td>
<td>1.20%</td>
</tr>
<tr>
<td>Rater Race Proportion</td>
<td>14.42%</td>
</tr>
<tr>
<td>Ratee Race Proportion</td>
<td>0.07%</td>
</tr>
<tr>
<td>Rating Bias x Validity</td>
<td>1.20%</td>
</tr>
<tr>
<td>Rating Bias x Rater Race Proportion</td>
<td>6.01%</td>
</tr>
<tr>
<td>Rating Bias x Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Validity x Rater Race Proportion</td>
<td>0.03%</td>
</tr>
<tr>
<td>Validity x Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rater Race Proportion x Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias x Validity x Rater Race Proportion</td>
<td>0.05%</td>
</tr>
<tr>
<td>Rating Bias x Validity x Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias x Rater Race Proportion x Ratee Race Proportion</td>
<td>0.06%</td>
</tr>
<tr>
<td>Validity x Rater Race Proportion x Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias x Validity x Rater Race Proportion x Ratee Race Proportion</td>
<td>0.05%</td>
</tr>
</tbody>
</table>
Table 4

Proportion of Variance Accounted for in Power Levels by the Main Effects and Interaction Effects of the Manipulated Parameters for a Large Sample Size (n = 2000)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude of Rating Bias (Rating Bias)</td>
<td>86.95%</td>
</tr>
<tr>
<td>Validity of Predictor Scores (Validity)</td>
<td>0.76%</td>
</tr>
<tr>
<td>Rater Race Proportion</td>
<td>9.92%</td>
</tr>
<tr>
<td>Ratee Race Proportion</td>
<td>0.06%</td>
</tr>
<tr>
<td>Rating Bias x Validity</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias x Rater Race Proportion</td>
<td>2.29%</td>
</tr>
<tr>
<td>Rating Bias x Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Validity x Rater Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Validity x Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rater Race Proportion x Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias x Validity x Rater Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias x Validity x Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias x Rater Race Proportion x Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Validity x Rater Race Proportion x Ratee Race Proportion</td>
<td>0.00%</td>
</tr>
<tr>
<td>Rating Bias x Validity x Rater Race Proportion x Ratee Race Proportion</td>
<td>0.03%</td>
</tr>
</tbody>
</table>
Table 5

*Power Levels for the Magnitude of Rating Bias X Rater Race Proportion Interaction*

<table>
<thead>
<tr>
<th>Rating Bias</th>
<th>90 / 10</th>
<th>80 / 20</th>
<th>70 / 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>0.10</td>
<td>0.26</td>
<td>0.48</td>
</tr>
<tr>
<td>0.30</td>
<td>0.15</td>
<td>0.41</td>
<td><strong>0.67</strong></td>
</tr>
<tr>
<td>0.45</td>
<td>0.18</td>
<td>0.50</td>
<td><strong>0.76</strong></td>
</tr>
</tbody>
</table>

*Note.* Power levels are collapsed over total sample size, validity of predictor scores and ratee race proportions. Bolded values indicate acceptable levels of power.
Table 6

*Power Levels for the Total Sample Size X Magnitude of Rating Bias X Rater Race Proportion Interaction*

<table>
<thead>
<tr>
<th>Rater Race Proportion</th>
<th>Total Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td>2000</td>
</tr>
<tr>
<td>Rating Bias</td>
<td>90 / 10 80 / 20 70 / 30</td>
</tr>
<tr>
<td>Rating Bias</td>
<td>90 / 10 80 / 20 70 / 30</td>
</tr>
<tr>
<td>Rating Bias</td>
<td>90 / 10 80 / 20 70 / 30</td>
</tr>
<tr>
<td>0.15</td>
<td>0.07 0.09 0.10</td>
</tr>
<tr>
<td>0.30</td>
<td>0.20 0.26 0.47</td>
</tr>
<tr>
<td>0.45</td>
<td>0.51 0.68 0.81</td>
</tr>
<tr>
<td></td>
<td>0.21 0.26</td>
</tr>
<tr>
<td></td>
<td>0.65 0.76</td>
</tr>
<tr>
<td></td>
<td>0.93 0.98</td>
</tr>
</tbody>
</table>

*Note.* Bolded values indicate power levels above acceptable levels (i.e., above 0.60 and 0.80). All power level values are collapsed over validity of predictor scores and ratee race proportion values. Rating Bias refers to the magnitude of rating bias parameter.
Table 7

*Power Levels Associated with the Combinations of Manipulated Parameter Values*

<table>
<thead>
<tr>
<th>Rating Bias</th>
<th>Validity</th>
<th>Total Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rater Race Proportion</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90/10 80/20 70/30</td>
</tr>
<tr>
<td>0.15</td>
<td>0.10</td>
<td>0.07 0.08 0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.07 0.09 0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.07 0.09 0.11</td>
</tr>
<tr>
<td>0.30</td>
<td>0.10</td>
<td>0.12 0.18 0.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.12 0.19 0.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.15 0.22 0.30</td>
</tr>
<tr>
<td>0.45</td>
<td>0.10</td>
<td>0.22 0.35 0.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.23 0.38 0.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.27 0.45 0.56</td>
</tr>
</tbody>
</table>

*Note.* Bolded values indicate power levels above acceptable levels (i.e., above 0.60 and 0.80). All power level values are collapsed over ratee race proportion values. Rating Bias refers to the magnitude of rating bias parameter. Validity refers to the validity of predictor scores parameter.
Figure Captions

Figure 1. Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.30, validity = 0.10.

Figure 2. Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.30, validity = 0.30.

Figure 3. Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.30, validity = 0.50.

Figure 4. Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.30, validity = 0.10.

Figure 5. Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.30, validity = 0.30.

Figure 6. Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.30, validity = 0.50.

Figure 7. Power results for rater race proportion and ratee race proportion combinations: total sample size = 500, magnitude of bias = 0.45, validity = 0.10.

Figure 8. Power results for rater race proportion and ratee race proportion combinations: total sample size = 500, magnitude of bias = 0.45, validity = 0.30.

Figure 9. Power results for rater race proportion and ratee race proportion combinations: total sample size = 500, magnitude of bias = 0.45, validity = 0.50.
Figure 10. Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.45, validity = 0.10.

Figure 11. Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.45, validity = 0.30.

Figure 12. Power results for rater race proportion and ratee race proportion combinations: total sample size = 1000, magnitude of bias = 0.45, validity = 0.50.

Figure 13. Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.45, validity = 0.10.

Figure 14. Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.45, validity = 0.30.

Figure 15. Power results for rater race proportion and ratee race proportion combinations: total sample size = 2000, magnitude of bias = 0.45, validity = 0.50.
Figures

Figure 1

![Graph showing the proportion of Rater Race \times Ratee Race terms significant out of 10,000 iterations for different Rater Race Proportions. The x-axis represents Ratee Race Proportions (70/30, 60/40, 50/50), and the y-axis shows the proportion of terms significant. The graph uses bars for each proportion, with legend indicating 50/10, 80/20, and 70/30.]
Figure 2
Figure 3
Figure 4

![Bar chart showing the proportion of ratee-ratee race terms significant out of 10,000 iterations for different ratee race proportions. The rates are indicated by the legend: 60/10, 80/20, and 70/30. The proportions range from 0.0360 to 1.0000.](image)

### Ratee Race Proportions

- 70/30
- 60/40
- 50/50
Figure 5
Figure 6

[Bar chart showing the proportion of ratee x ratee race terms significant out of 10,000 iterations for different ratee race proportions: 70/30, 60/40, 50/50. Each proportion has three bars representing different rater race proportions: 60/10, 80/20, 70/30.]

Differential Prediction 59
Figure 7
Figure 8

![Graph showing the proportion of Rater Race x Ratee Race terms significant out of 10,000 iterations.](image)

- **Y-axis**: Proportion of Rater Race x Ratee Race terms significant out of 10,000 iterations.
- **X-axis**: Ratee Race Proportions.
- **Legend**:
  - 60/40
  - 80/20
  - 70/30

The graph illustrates the relationship between ratee race proportions and the proportion of significant terms across different raters' race proportions.
Figure 9
Figure 10

The diagram shows the mean proportion of ratee race x ratee race terms significant out of 1000 iterations. The X-axis represents the ratee race proportions (70/30, 60/40, 50/50), while the Y-axis shows the mean proportion. The bars are color-coded to represent different rater race proportions: 60/10 (light gray), 80/20 (gray), and 70/30 (dark gray).
Figure 11

[Bar chart showing the mean proportion of rater race x ratee race terms significant out of 1000 iterations for different ratee race proportions (70/30, 60/40, 50/50)]

Legend:
- 60/10
- 80/20
- 70/30
Figure 12
Figure 13

[Bar chart showing the proportion of ratee race x ratee race terms significant out of 10,000 iterations for different ratee race proportions.]

- **Ratee Race Proportions:**
  - 70/30
  - 60/40
  - 50/50

- **Legend:**
  - 60/10
  - 80/20
  - 70/30
Figure 14

![Bar chart](image)

- **Y-axis:** Proportion of Rate x Rate x Rate terms significant out of 100,000 iterations
- **X-axis:** Ratee Race Proportions (70/30, 60/40, 50/50)
- **Legend:**
  - 60/10
  - 80/20
  - 70/30

Differential Prediction 67
Figure 15