2007-01-01

Wherry Revisited: An Empirical Examination of the Nonperformance Factors that Influence Variation in a Performance Rating

Raymond B. Roll EdD
University of San Diego

WHERRY REVISITED: AN EMPIRICAL EXAMINATION OF THE NONPERFORMANCE FACTORS THAT INFLUENCE VARIATION IN A PERFORMANCE RATING

by

RAYMOND B. ROLL

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Education
University of San Diego

January 2007

Dissertation Committee

Fred J. Galloway, Ed.D., Chair
Port R. Martin, Ed.D., Member
Theresa M. Monroe, Ed.D., Member

Abstract

In a perfect performance rating system, both the recall and rating of an individual's behavior would precisely mirror the performance of that ratee. However, the reality of performance rating systems is that often times the rater's recall and subsequent rating fails to reflect the true performance of the individual. The difference between actual and perceived performance has been attributed in the literature to conscious or unconscious rater bias.

In 1952, Wherry developed a rating theory based on a series of mathematical equations that precisely defined the relationship between the performance of the ratee and the recall of that observation. Key to his theoretical work was the fundamental rating equation, which stated that a rating score was equal to the actual performance of the ratee plus an observation and recall bias component as well as random error. As such, the goal of this study was to test the appropriateness of this framework by applying it to an actual performance rating system used by the United States Navy on board a particular ship. By utilizing Wherry's basic theory, together with data on rater and ratee nonperformance characteristics (e.g. gender, race, education, height, smoker/non-smoker, etc.), multiple regression analysis was used to identify the nonperformance factors that affected the accuracy of a rating process for 423 individuals.

The results of this study supported Wherry's theory in that four of the eight variables contained in the study's final regression model strongly indicated the existence of rater bias. Ratees that were either white, had personality types that matched the first raters, or were of the same race as the second raters generally received higher evaluation scores than ratees that were not, while ratees that smoked received lower evaluation

scores. Even though more research is clearly needed to determine the factors that may have produced these biases, their existence in such a high-stakes performance appraisal system suggests that at a minimum, the Navy needs to develop a strategy that educates its raters on the possibility that they might be subconsciously discriminating against others based on their race, personality match, and smoking preference.

DEDICATION

This work is dedicated to my wife, Marie, and my three loving children: Peter, Jacob and Alexandra, whose love and understanding were great enablers. Their unwavering patience and support were instrumental in the completion of this dissertation.

ACKNOWLEDGEMENTS

I would like to thank my immediate family for their extreme patience during this dissertation process. Their love, support, and understanding inspired me throughout this long journey. A special "thank you" goes to my wife, Marie, who took on many extra burdens that freed me to concentrate on this study. Peter, Jacob, and Alexandra, thank you for being the great children that you are.

My deepest gratitude and appreciation goes to my committee for their encouragement and wisdom. Dr. Fred Galloway's sage advice and guidance were vital in completing this work. Dr. Theresa Monroe's mentorship was crucial in understanding the underlying currents of this study and Dr. Bob Martin's steadiness through troubled times righted the ship and get me on course.

To my Mother, Father, brothers and sisters, thank you for your understanding of my absence from the family during this dissertation timeframe, I hope to re-engage soon. And finally, I would like to thank all my friends who have provided much encouragement and warm support – thank you all.

TABLE OF CONTENTS

LIST OF TABLES

Background

Performance appraisals are widely used and serve a number of important functions within organizations. Each year in the United States over 70 million individuals receive some type of performance appraisal (Matens, 1999). A vast majority of these appraisals will consist of a performance rating system (Landy & Farr, 1980). This type of system requires raters to use their judgment, based on past observations, to measure and then rate an individual's performance according to a scaled rating arrangement. The results of this rating process are then used as a basis for many personnel decisions, including salary increases, recommendations for promotion, transfer, release, or training programs, as well as for ratee development and performance feedback (Cleveland, Murphy, & Williams, 1989).

Considering the number of annual performance ratings conducted and the important roles the results of these ratings play within organizations, it is not surprising that rating accuracy is a primary concern of performance appraisal research. Despite being studied for over eight decades, there has been a consistent dissatisfaction in the literature with rating accuracy on the part of both researcher and practitioner (Landy & Farr, 1980). Research has shown that only 20 percent of all appraisals are considered effective in assessing work performance (Matens, 1999). As a result, a significant portion of the existing research has examined the factors that contribute to the overall effectiveness of performance appraisals (Keeping & Levy, 2000).

Achieving measurable improvements in performance effectiveness has proven to be difficult. This is due, in part, to the inherently biased nature of performance appraisals. Personal judgments about an individual's performance are inescapable. Subjective values

and fallible human perceptions are essential to the process (Oberg, 1999). Additionally, "researchers in the field of decision-making have long been aware of and have studied systematic biases in human judgment that represent deviations from a rational model" (Schoorman, 1988). These biases and judgmental errors introduce measurement error in the assessment of performance and, of course, directly affect the accuracy and the effectiveness of the ratings.

In an effort to mitigate these inherent and systematic problems, a substantial amount of research has been conducted in an attempt to improve the "validity of judgmental measurements of performance" (Landy & Farr, 1980). The goal of the majority of this research was to determine "what factors other than actual performance of the ratee affect performance ratings and to determine methods by which these biases could be eliminated or minimized" (Wherry & Bartlett, 1982). Throughout the years researchers have attempted to increase the effectiveness of performance appraisals by improving rating format, designing techniques to improve long-term rater recall, or by developing training programs to aid the rater's recall of ratee performance (Borman, 1979).

In his comprehensive review of the literature, R. J. Wherry, proposed that the accuracy of a rating hinges on three chronological steps: the performance of the ratee, the observation of this performance by the rater, and the recall of this observation by the rater (Wherry & Bartlett, 1982). Wherry reasoned that the rating's accuracy is directly affected by biases that enter into the performance rating process during both the perception of the observed behavior and during the recall of this behavior (Wherry & Bartlett, 1982). He

theorized that these biases could either be positive towards the ratee or negative against the ratee.

In 1952, Wherry developed a rating process theory that defined in mathematical terms the relationships between the performance of the ratee, the observation of that performance by the rater, and the recall of that observation. The fundamental rating equation of his theory stated that a rating score is equal to the actual performance of the ratee plus a rater observation and recall bias component plus a random error term (Wherry & Bartlett, 1982). In order to further break down the overall process, Wherry proposed that the performance, observation, and recall components were each made up of a systematic portion and a random portion.

According to Wherry's theory, the systematic portion of the performance of the ratee component was determined to be a function of the ratee's true ability and the influence of the work environment. Examples of this work environment factor were the training provided to the ratee, tools used to perform the task, and the work setting. Lighting conditions, temperature conditions, and noise levels were examples of the work setting (Wherry & Bartlett, 1982). The rater observation component was determined to be a function of the performance of the ratee and a bias of observation. This bias factor was described as a "bias of perception" that would vary in magnitude depending on the number of relevant contacts that the rater had with the ratee (Wherry & Bartlett, 1982). Lastly, the rater recall component was determined to be a function of all the rater's observations of the ratee and a bias of recall. For Wherry's theory, the bias of recall and the bias of perception were "assumed to follow a general pattern where inconsistent

details were obliterated in favor of a general concept, while supporting detail was selected or even unknowingly invented" (Wherry & Bartlett, 1982).

The various factors that made up the systematic portion of each of these components were then defined by a series of linear equations that were substituted back into Wherry's fundamental rating equation. The three rating components' random portions were also defined mathematically and inserted into Wherry's fundamental rating equation as a series of random error terms. The resultant rating equation is rather long and complex, reflecting the complexity of an actual rating process. Wherry argued that the linear equations contained in this final rating equation were testable and provided a reliable method of measuring rating bias.

Wherry hoped that his theoretical formulations would encourage further research that would lead to a better understanding of the magnitude and source of rating bias. He believed this understanding would enhance the ability to control this bias and enable researchers to accurately measure improvements in rating effectiveness (Wherry & Bartlett, 1982). However, after fifty years of performance rating research since Wherry's proposal, these biases still plague performance rating systems.

Problem Statement

In a perfect performance rating system, the recall and the rating of the observed ratee behavior would mirror the true performance of that ratee. However, the reality of performance rating systems is that rater recall does not always equal ratee performance. The one universally accepted finding of all the research on performance rating systems is that the ratings are often plagued by a host of problems, including halo and leniency tendencies, unintentional manipulation, and race, gender, or age biases (Facteau & Craig,

2001). These often unconscious biases adversely affect the ability of appraisal systems to accurately assess ratee performance and induce unwanted variation in performance ratings.

Wherry argued that in order to improve rating accuracy it was paramount to identify and control the biases that occurred during the observation and the recall of ratee performance. To achieve this, Wherry presented a performance rating theory based on a series of testable linear equations. He believed that once these biases were isolated and understood, appropriate methods could then be developed to improve the assessment capabilities of performance appraisals. In their review of rating research, Landy and Farr (1980) encouraged performance rating researchers to closely examine and empirically test Wherry's performance rating theory. Regrettably, Wherry's theories on the rating process have gone virtually unacknowledged among performance rating researchers and little follow up research has attempted to validate or build upon his theories (Wherry & Bartlett, 1982). Additional empirically based research is needed to understand the magnitude of the biases that influence the rater's observation and recall of ratee performances.

Purpose of the Study

The goal of this study was to take Wherry's theoretical rating framework and apply it to an actual performance rating system used by the US Navy. By utilizing his basic theory, together with data on rater and ratee nonperformance characteristics (e.g. gender, race, education, height, etc.) this study used multiple regression analysis to quantify the nonperformance factors that affected the accuracy of an actual rating process for 423 US Navy sailors. In addition to empirically testing Wherry's framework, this

study identified the extent and direction of the bias inherent in this particular performance rating cycle.

The primary goal of Wherry's rating theory was to provide a method to identify and then reduce variation in order to improve rating accuracy. Wherry hoped his linear equations would help isolate and then control the biases that enter into the rating process. Controlling the variation induced by biases remains a high priority. If performance ratings are not accurate and do not truly reflect ratee performance then their use as a tool for basing personnel decisions is questionable if not unjustifiable.

The ultimate goal of this study was to enhance the overall understanding of the performance rating portion of the appraisal process. This increased understanding may lead to research designed to improve rating formats or lead to the development of improved rater training that will lead to more effective appraisal systems.

Research Questions

This study was designed to answer the following research questions:

1. What are the statistically significant, nonperformance factors that influence variation in the rating scores of an actual performance appraisal process?
2. Is there a specific rater bias that can be attributed to individuals displaying "like" characteristics as the rater as opposed to individuals that display different characteristics?
3. Can these statistically significant, nonperformance factors be attributed to rater bias and can they then be used to support Wherry's theoretical performance rating equation?

Literature Review

Introduction

The majority of the studies conducted on performance rating systems can be grouped into three broad research areas: format, cognitive processes, and rater/ratee characteristics. Early research focused primarily on format in an attempt to increase the accuracy of performance appraisals by improving the vehicle used for the ratings. Later research has been dedicated to rater cognitive processes in the hopes of designing techniques to improve long-term memory recall or to develop improved training programs designed to aid raters in improving their observation and recall of ratee performance (Borman, 1979). The third area of research has concentrated on rater/ratee characteristics to determine if the interactions between rater and ratee resulted in measurable biases.

As might be expected, not all studies on performance ratings readily fall into these three broad areas. For example, there is considerable research on the effect the actual ratings have on both the ratee and the rater. However, studies that concentrated on ratee reaction are not covered by this review since they do not directly add to the literature on improving accuracy. In addition, research on how raters or ratees can manipulate the performance rating system was also not included. This paper is based on the assumption that both raters and ratees are acting in a forthright way and not purposely attempting to manipulate the rating process in one direction or the other.

Three extensive reviews of the research on performance ratings have previously been conducted. R. J. Wherry's review covered the research conducted on rating systems prior to 1950, and Frank Landy and James Farr's review covered the research performed

between 1950 and 1980. Wherry's research was conducted to support a US Army study and his findings were not widely published. As a result, his work has received relatively little attention among performance appraisal researchers. However, Landy and Farr's research was exhaustive and is widely praised within the performance appraisal field for its thoroughness. In addition, Borman, Buck, Hanson, Motowidlo, Stark, and Drasgow's (2001) review of rating format research effectively covered the time period following Landy and Farr's review up to 2001. The review of the literature that follows will not repeat the findings of these three reviews but only cite pertinent passages as they relate to the research within the three basic research areas. The following sections will provide an overview of the research conducted within these three areas and comment on the limitations to this existing research.

Format Research

Due to the seemingly limitless variety of performance ratings used by organizations, the research on rating format has been extensive. Format, as it applies to performance appraisals, is the "physical arrangement in which the rating-scale definition and levels are presented to the rater for application to stimuli" (Madden & Bourdon, 1964). Researchers believe that format should aid raters by assisting their recall of ratee performance in an efficient and organized way. They also believed that format should help raters translate their recall of ratee behavior into information relevant for making accurate evaluation judgments (Borman et al., 2001). Researchers have proposed that since the rating scale's format is the vehicle by which a rater makes and communicates his evaluation judgments, its importance cannot be overemphasized (Madden & Bourdon,

1964). This perception has helped drive the desire to develop the most effective rating format.

Landy & Farr (1980) believed that an ideal performance measurement should include a combination of objective, personal, and judgmental values. However, due to the difficulty in applying objective and personal information across different individuals and tasks, judgmental rating scales have become the primary tool for performance appraisals. Rating scales were first introduced to the general psychological community in 1922 and freed raters from making quantitative based judgments to appraise the ratees' performances (Landy & Farr, 1980). Rating scales enabled raters to make as fine a distinction in their ratings as they desired.

In the early, developmental years of rating scales, format manipulations were incremental with slight improvements or adjustments being made to previous rating formats. An example of this incremental approach is represented by Madden and Bourdon (1964). They researched the effect on rating reliability with regard to the physical placement and style of the rating scale used to record the ratings (either horizontal or vertical, bars or no bars, or with various numbering methods.) They found that there was a difference in judgment and preference that could be attributed to the format of the rating scale but made no effort to determine which format was optimal. Later studies showed that raters may have a preference for specific formats but these preferences had little effect on actual rating accuracy (Landy & Farr, 1980).

Borman (1977) noted that early research resulted in the development a number of rating scales and cited graphic rating scales, forced-choice formats, man-to-man rating scales, and forced-distribution formats as the most widely used. As these formats were

developing and becoming more judgmental and less quantitative, researchers and users began to look at the foundations of these rating systems with greater skepticism. Wherry's review of the literature during this time period revealed that most rating systems were based on "an abundance of platitudes and rules-of-thumbs, a smattering of empirical findings, and a complete absence of any rational system or theory" (Wherry & Bartlett, 1982).

To offset this skepticism, subsequent research efforts attempted to evaluate the effectiveness of these early rating scales in recording valid performance information. Unfortunately even after extensive research was expended, no clear guidance as to which scale was best has ever resulted from this research (Borman, 1979). As rating formats continued to develop away from quantitative based judgments, the disenchantment by users and researchers in the subjective and arbitrary nature of rating systems grew (Landy & Farr, 1980). Inaccuracies within the systems began to negate the usefulness of the performance ratings.

Responding to the need to improve performance rating accuracy, P. C. Smith and L. M. Kendall proposed a format in 1962 that was considered a significant advancement. They built upon the "critical incidence" notion introduced by J. C. Flanagan in 1954 by adding behavioral expectations scales, later to be transformed into Behaviorally Anchored Rating Scales (BARS) (Borman et al, 2001). BARS were added to the rating scales to serve as anchors to help raters make more accurate judgments in their ratings. This new concept captured the energies of researchers and has since dominated rating system development efforts.

In 1972, R. Blanz and E. E. Ghiselli advanced the BARS approach by adding a mixed standard scale. This format had an effective, mid-level, and ineffective behavioral statement for each dimension being rated. The rater indicated whether the ratee was more, less, or at the same level for each one of the behavioral statements (Borman et al., 2001). A later variation of the BARS required the rater to make a judgment on the frequency that the ratee exhibited a specific behavioral statement (Borman et al., 2001).

Additional studies on the effectiveness of BARS based rating systems found that raters often had difficulty discerning behavioral similarities between actual ratee performances and the sometimes very specific behaviors used to anchor the scales (Borman, 1979). In an effort to eliminate these difficulties Borman developed his variant of the BARS. He suggested using more general anchors with a wider range of described behaviors for each dimension being rated. He hypothesized that more general descriptions would increase the probability of raters matching observed behaviors with the scaled behavior (Borman, 1979). In 1986, J. S. Kane introduced yet another version of BARS by including a "negative range avoidance score" which attempted to measure how well the ratee avoided ineffectual performance (Borman et al., 2001).

Recently, research has attempted to capitalize on the extensive use of computers within organizations. Borman et al. (2001) designed a study utilizing a Computerized Adaptive Rating Scale (CARS) format. This scale presented the rater with a series of paired behavior statements to compare against observed ratee behavior. Based on the rater's selection of which of the paired statements best described the ratee's behavior, another set of behavior statements would be presented to the rater. Each dimension being rated would have its own series of paired behavioral statements. They reasoned that the

responses to each successive paired statement would provide a more precise estimate of actual ratee performance (Borman et al., 2001). Their study compared the results of the CARS format against a graphic rating scale format and the BARS rating format. They found that the CARS format resulted in significantly higher accuracy and validity than these other scales (Borman et al., 2001). However, even with these favorable results these researchers only suggested that the CARS format might be an effective method for some applications.

Limitations of the Research on Format

The research on rating format has been driven by the belief that raters could be aided by format in their recall of ratee past behavior and that advances in rating format would increase the accuracy of performance appraisals. However, measurable gains in accuracy have not been delivered to actual performance ratings within actual organizations. Format comparison studies have generally shown small differences between formats in terms of the level of rater errors, reliability, validity, or accuracy (Landy & Farr, 1980; Borman et al., 2001).

This is not to say that the rating formats used today by organizations are not more accurate than systems used in the past. However, it has been difficult to quantify any real improvements in accuracy. Borman (1979) believes that the inability to quantify improvements in accuracy primarily stems from the difficulty in establishing a definitive "true score" against which to compare format improvements. This difficulty can be attributed to the subjectivity of raters as to what behaviors indicate good, bad, or standard performance (Smith & Kendall, 1962). Without agreement on a standard behavior in which to use as a benchmark, the ability to quantify the degree of accuracy among

various formats may be unattainable and slows the adoption and acceptance of format improvements.

A small percentage of the research on format has concentrated on designing training programs to aid the raters in understanding and using rating systems more effectively. The research on format training has also not produced measurable improvements in reducing rater errors. Borman's review (1979) of the literature covering the studies on format training found that only some of the training programs appear to be successful in reducing certain rating errors while other rater errors persisted or were even exacerbated. Again, Borman argues that most studies were unable to produce a viable true score to compare the trainees' ratings against (Borman, 1979). This inability to accurately measure gains in format training programs has hampered the acceptance of research based training improvements for use in actual organizations.

Cognitive Process Research

After F. S. Landy and J. L. Farr (1980) completed their review of the research on format, they called for a moratorium on future format research. They had estimated that appraisal format accounts for less than six percent of appraisal accuracy and stressed that research in other areas was needed. They recommended that significantly more research was required on the rater's mental processes as they pertained to performance appraisals. "We must learn more about the way in which potential raters observe, encode, store, retrieve, and record performance information, if we hope to increase the validity of ratings (Landy & Farr, 1980).

The research in cognitive processes began in earnest after this recommendation. Researchers exploring the rater's cognitive processes believe that in order to improve

rating accuracy you must first understand the rater's decision-making process. Cognitive process researchers believe that once these processes are understood, training programs, information storage techniques, and/or format changes could be developed to increase performance rating effectiveness by reducing the variation among raters.

A significant portion of this research has concentrated on memory encoding and on recall accuracy. Researchers reasoned that if raters are to provide accurate ratings, they must be able to reliably store and then access performance information stored in their memory. DeNisi & Peters (1996) suggested that "a rater's ability to accurately recall information is largely dependent on how well the information was organized in memory during the encoding process." They attempted to test this theory in a field setting and designed a study to determine whether structured diary keeping and structured recall affected the recall of performance information. They found that raters who kept diaries produced ratings that were less elevated and were able to discriminate better both within, and between ratees, than those that did not. They also found that organizing performance information through very structured diary keeping had a positive effect on recall and ratings (DeNisi & Peters, 1996).

In another study, Cafferty, DeNisi, and Williams (1986) found that raters primarily acquired information either grouped by persons (one ratee performing different tasks), grouped by tasks (multiple ratees performing the same task), or in an ungrouped fashion across both raters and tasks. They found that raters that grouped information by person or by task resulted in more accurate recall and thus more accurate ratings than those that acquired information in an ungrouped manner (DeNisi & Peters, 1996).

K. R. Murphy and W. K. Balzer (1986) have "argued that under certain conditions, raters may depend more on their general impressions of ratees than on their memory of specific details." They designed a study on long-term memory recall to test this theory. The results of that study supported their argument and seemed to indicate that the reliance on general impressions did not necessarily mean rating accuracy decreased. In their study, ratings were more accurate for long-term, impression-based recall than for short-term, immediate recall ratings (Murphy & Balzar, 1986). It is their hypothesis that being able to reliably assess an individual's overall performance may be more important for appraisals then being able to remember the "subtle nuances of behavior" (Murphy & Balzar, 1986).

One study, not directly related to appraisal memory encoding and recall, offers yet another method that raters may use to remember rating information. C. A. Hamilos and G. F. Pitz (1977) designed a recognition test to explore an individual's encoding of quantitative information. A portion of their study was designed to determine if the subjects could discriminate between new data and data that they had seen previously. They found that subjects were able to discriminate the old data from the new more effectively when the data presented to them was on the extreme minimum or maximum values of the old data. Their findings suggest that there is a possibility that raters may use the ratee's extreme behaviors rather than their standard behaviors as a basis for making rating judgments.

One segment of the cognitive process that requires more research concerns the rater's perception of the organization's appraisal environment and how this perception influences the rater's appraisals (Cleveland, Murphy, & Williams, 1989). Performance

ratings can serve several purposes within organizations, and the rater's perception of the performance rating's primary purpose can have a significant impact on how appraisal judgments are made (Cleveland, Murphy, & Williams, 1989; Landy & Farr, 1980). However, little research has attempted to measure the rater's understanding of the organization's intended appraisal use and how that understanding affects ratings effectiveness. Appraisals that provide feedback are significantly different from appraisals that serve as a guide for making personnel decisions. Studies have shown that appraisals conducted for feedback or for developmental purposes are less prone to rating bias than are appraisals that are conducted for administrative decision-making purposes (Williams, DeNisi, Blencoe, and Cafferety, 1985). Williams et al. (1985) cite the findings of studies conducted by Fisher and McGregor in suggesting that raters dislike giving poor ratings in general. They went on to say that further studies have shown this aversion to giving poor ratings is increased if the rater knows that the ratings will be viewed by the ratees (Williams et al., 1985). Wherry and Bartlett (1982) found that if the rater knows that the rating will have to be justified to the ratee then the rater may have a tendency to recall a greater number of favorable perspectives leading to higher leniency in the ratings.

Rater performance recall may also be affected by previous interaction between rater and ratee. Hogan (1983) found that there was a significant positive relationship between initial expectations and later performance evaluations. Schoorman (1988) found that supervisors who positively participated in the hiring of individuals gave higher evaluations and promotion recommendations than did those who not participate in hiring decisions. Additional research has found that previous ratings given by a rater serve as an anchor for future ratings and may increase halo or leniency inaccuracies.

Limitations of the Research on Cognitive Processes

The most common criticism lobbied against the research on cognitive processes has been that most of the research was conducted in a laboratory setting where important process issues that occur in real organizations are not present (DeNisi & Peters, 1996). Issues concerned with short-term (lab setting) versus long-term (organizational) memory recall and the recollection of one event (lab setting) versus a multitude of events (organizational) are cited most frequently. It is believed that additional field research is needed in the cognitive process area before any generalized findings can be presented. Research continues in the hope that significant findings in cognitive processes could have a major impact on format development and rater training programs.

Rater/Ratee Characteristics Research

The third broad area of research has focused on analyzing performance ratings as a function of rater and ratee demographic characteristics. This research has attempted to isolate inaccuracies or variation caused by nonperformance factors. The majority of this research has centered on how gender, race, or age bias affected rating accuracy (Hartel, Douthitt, Hartel & Douthitt, 1999; Landy & Farr, 1980; Bigoness, 1976; Hamner, Kim, Baird & Bigoness, 1974; Pulakos, White, Oppler, & Borman, 1989; Nevill, Stephenson & Philbrick, 1983). These studies have resulted in few universal findings of significance considering the magnitude of the research. Landy and Farr (1980) believe that these researchers have too narrowly focused their studies by looking at too few demographic characteristics or by just concentrating on either the rater or ratee characteristics singularly. They and other critics believe that unmeasured or hidden variables may have had an unknown effect on the results of the studies (Landy & Farr, 1980).

When the interaction between rater and ratee demographic characteristics has been studied, it has generally been limited to the effects of race or gender (Mobley, 1982; Pulakos & Wexley, 1983; Schmitt & Lappin, 1980; Landy & Farr, 1980). The results of these studies do seem to indicate that there is a positive relationship when rater and ratee race are similar. The results of rater and ratee gender research are more mixed but there does seem to be indications that male raters rate female performances lower than they do males (Landy & Farr, 1980).

One segment of rater and ratee characteristics research that has recently gained increased attention focuses on rater and ratee personality traits. Employee personality traits have been used by managers as estimates for potential performance during personnel selection purposes, but little research has been done to examine the effect that rater or ratee personality traits have on performance ratings (Borman, Hanson, & Hedge, 1997). Most of the limited research that has been done in this area seems to suggest that personality has little influence on actual performance ratings (Lefkowitz, 2000).

The research on what effect characteristic similarities shared by the rater and the ratee have on performance appraisals is also limited. Kirsch & Zalesny (1986) findings indicated that rater/ratee differences in specific characteristics might have an even greater effect on ratings than the effect of being similar. Others have found that when raters perceive there are similarities between themselves and the ratees then that perception has an even greater effect on performance ratings than the existence of actual similarities (Strauss, Barrick, & Connerly, 2001; Turban & Jones, 1988). Additional study needs to be conducted to verify these studies, but these early findings indicate a possible source of rater bias in performance ratings.

Limitations to Existing Research

As the previous sections indicate, research on format, cognitive processes, and rater/ratee characteristics has been extensive. However, little of the research in these three areas seems to have influenced the performance appraisal process utilized by most managers. Performance rating systems still suffer from inaccuracy and variation. This is, in part, due to the previous cited criticism of this research. These and other limitations have significantly restricted the applicability of performance appraisal research in an organizational setting.

A major limitation to a majority of the research has been the difficulty in transferring findings found in a laboratory setting into real world organizations (Bernardin & Villanova, 1986). Most laboratory studies have followed traditional research design by holding all things constant with the exception of the focus factor of the study. It has been found that the results from this type of study are influenced by the research design and are not readily applicable to actual performance appraisal processes found in real organizations (Wendelken & Inn, 1981). Results from laboratory studies that focus on staged events and the subsequent rating of the information observed during these events, in isolation from all external factors, do not translate well to a manager that must sort through a multitude of internal and external stimuli over an extended period of time (Landy & Farr, 1980).

Even when research has been conducted in a field setting there are limitations to the findings of the study. Although Wherry was able to reduce the performance rating process into three basic steps the actual process is much more complex. Landy and Farr (1980) identified thirteen components for their proposed model of a performance rating

process. This complexity is magnified by the many different characteristics of organizations that utilize these systems, the number of uses for performance ratings within these organizations, and the virtually limitless variety of rating formats designed to support these uses (Landy & Farr, 1980). Additionally, every organization has unique set of internal and external influences that frame or shape the organization's culture. "Each organization has a different idea of what may be important in assessing their people; consequently, each rating instrument is ultimately unique" (Landy & Farr, 1980). Due to this uniqueness and the variety of rating formats, research gained from one organization may not be applicable to other organizations.

A major limitation to the existing research on performance rating systems has been its inability to accurately measure improvements. Without the ability to quantify the improvement gained by a change in a rating system, it has proven difficult to provide the concrete numbers that managers require before they will embrace the new system (Borman 1979). Therefore, advances in performance rating systems have been slow to be adopted by practitioners.

Summary

Past research efforts in these three board areas: format, cognitive processes, and rater and ratee characteristics, have all attempted to improve rating accuracy. Researchers have tried to find ways to reduce the bias within rating systems to improve the effectiveness of the performance rating. Format research attempted to decrease bias by aiding rater recall of past observations and by providing behavior statements to lesson recall biases. Format research has also tried to reduce arbitrary judgmental errors by providing anchors to base rating judgments. Cognitive processes research concentrated

on reducing biases induced during the observation of ratee behaviors and during the subsequent recall of those observations. And rater/ratee characteristics research attempted to identify the sources of bias in the hopes that once identified, the biases could then be eliminated.

Wherry & Bartlett (1982) wrote, "If the perceiver makes a conscious effort to be objective, after becoming aware of the biasing influence of previous set, he may be able to reduce the influence of his bias." However, without solid empirical proof that rater biases exists and they do in fact adversely affect performance ratings, raters have little motivation to change or reduce the influence of their biases.

Methodology

Introduction

As previously discussed, conscious or unconscious biases can induce unwanted variation in performance ratings. Although much empirical work remains to be done to fully understand the source and magnitude of these biases, R. J. Wherry provided a theoretical rating framework designed to recognize and reduce bias-based variation in ratings. Building on this framework, this study quantified the nonperformance factors that affected the accuracy of an actual performance rating system. Quantitative research methods were utilized to accomplish this analysis. This analytical technique captured the measurable nonperformance factors of a specific performance rating process. Data for the study came from a real organization obtained during an actual performance rating appraisal and were used to construct three "progressive" multiple regression

models. Each of theses models were specifically designed to answer the following research questions:

1. What are the statistically significant, measurable nonperformance factors that influence variation among raters in an actual performance appraisal process?
2. Is there a specific rater bias that can be attributed to individuals displaying "like" characteristics as the rater as opposed to individuals that display different characteristics?
3. Can these statistically significant, nonperformance factors be attributed to rater bias and can they then be used to support Wherry's theoretical performance rating equation?

Previous studies on the effect of nonperformance factors on rating appraisals have generally examined only one or a few variables at a time. The results of these studies have been questioned because of the probability that additional factors that were not controlled by the study had an unmeasured effect on the study's results. This study attempted to avoid this criticism by including as many measurable nonperformance variables as possible. Each of these variables will be discussed within this chapter, as well as, each of the three regression models that were utilized.

Participants and Data Collection

The participants for this study consisted of individuals that made up the three lowest enlisted rankings (and their associated raters) that were evaluated during one performance rating cycle on board a US Navy aircraft carrier. This group was chosen because it offers a large sample size of ratees and raters. This data set also contained a

variety of rater to ratee combinations (from raters that rated a total of two ratees to raters that rated up to 50 ratees.)

The US Navy's rating system uses a two rater process to evaluate each ratee. Each ratee was initially rated by his immediate supervisor or the first rater. Each of the first raters rated only the ratees that were assigned to his or her division or work unit. The results of this rating process were then reviewed by a second rater who was senior to both the ratee and the rater. The second rater reviewed the ratings of all the first raters that were assigned to his or her work unit. The second rater could adjust the evaluations of the first rater if desired or leave the ratings as they were. The ratee was then given one final rating that was agreed to by both raters.

In addition to the evaluation results, this study also used rater and ratee demographic data that the Navy had on file and was supplemented by a personality profiler that estimated the personality type for each rater and ratee. The Navy maintains a personal file on each of its members called their "service jacket." Most of the on-file data was collected directly from the service jackets of each of the ratees and raters. The evaluation scores or performance ratings for all the rated individuals during the evaluation period were also collected directly from the ratees' service jackets.

For each evaluation the ratee was rated on seven performance characteristics based on a 0 to 5 scale with 0 being extremely poor performance and 5 being the best. Those seven scores were then averaged to come up with an overall performance rating score that ranged from 0.0 to 5.0. The overall evaluation average score was recorded for use in the regression models as the dependent variable.

A personality profiler was used to determine a measure for the raters' and ratees' personality types. Every evaluated ratee and each rater was given a profiler. This profiler was obtained from Human Resource Dimensions, Inc. and was developed by Donald A. Johnson, PhD. The development of the profiler was based on the research of Carl Jung, who suggested that differences in personality can be attributed to behavioral preferences. This profiler was chosen because it relies on behavioral preferences, and it is a statistically validated short form of a larger personality profiler used by the company. An individual's personality was profiled from four "perspectives." Each perspective compared personality preferences, based on the responses of a series of 48 paired questions, and profiled an individual as one or the other of the following four pairs: extroverting or introverting (E/I), sensing or intuiting (S/N), thinking or feeling (T/F), and organizing or adapting (Z/A). As an example, a person could be profiled as introverting, sensing, thinking and adapting (ISTA.) The personality profiles of the raters and ratees were included in the models to see if their interactions had an influence on performance appraisal variation.

One variable that was not readily available on record but was desired for the models was a measure of whether the rater and/or the ratee smoked cigarettes. This variable was desired to test if there was a bias for or against smokers. In order to measure whether an individual smokes or not, one question was added to the Identification Section of the personality profiler to ascertain the rater/ratee preference on smoking.

Analytical Methods

The goal of this analysis was to quantify the measurable nonperformance factors that influence the accuracy of this performance rating cycle. The quantitative data

collected was used to construct three multiple regression models that progressively built upon the previous models. Progressive multiple regression modeling was selected to allow the results of each model to be examined and compared against the results of the previous models. The first model attempted to isolate the effects of ratee demographic characteristics on the variation in performance rating scores. The second model examined the effect of matched rater and ratee demographic data to see if their interaction influenced the variation in performance scores. The statistically significant variables that were found to be common to both models were then closely examined to see if their influence on variation changed.

Independent Variables and Category Breakdown

For each of the three models the independent variables were made up of the statistically significant demographic and personality characteristics. The following independent variables were available for use in the three models. (Parentheses indicate source of the data, either on file in Navy records or through the personality profiler):

> Race (On file), Age (On file), Gender (On file), Height (On file), Weight (On file), Education level (On file), Number of months assigned to the ship (On file), Standard entry test scores (On file), Home of record (On file), Discipline record (On file), Personality type (Profiler), Married (On file), Number of children (On file), Smoker/Nonsmoker (Profiler)

These variables were broken down into the following categories for inclusion in the models:

> Race – Asian, Black, American Indian/Alaskan, White, Other
> Gender – Male or female
> Education Level – Less than high school degree, high school degree, some college, college degree
> Home of record – Northeast, Southeast, Mid-West, North Plains, South Plains, Northwest, Southwest, or Outside the US
> Discipline record –NJP or no NJPs.
> Children – Children or no children

Married – Married or not married
Personality type – (As determined by personality profiler)
Smoker – Smoker or nonsmoker

Additionally, the following continuous data was broken down into the following categories: less than one standard deviation below the mean, within one standard deviation below the mean, within one standard deviation above the mean, or more than one standard deviation above the mean:

Age, Months Assigned to the Ship, Standard Entry Test Scores, Height, and Weight

For model number One, these variables were inserted into the model first as continuous data and then in these categorical groups as dummy variables. This was done to capture the effect of these variables on evaluation scores both as continuous data and categorical data.

Justification for the Independent Variables

The demographic variables race and gender have been the subject of many performance appraisal studies (Hartel, Douthitt, Hartel & Douthitt, 1999; Landy & Farr, 1980; Bigoness, 1976; Hamner, Kim, Baird & Bigoness, 1974; Pulakos, White, Oppler, & Borman, 1989; Nevill, Stephenson & Philbrick, 1983). From these studies, there was sufficient evidence to believe that race and gender would have an effect on performance ratings. The Age variable was included in this study to examine whether maturity level had an impact on performance level. Height was included in the model to test the hypothesis that up to a point, taller male individuals are given higher evaluation scores than shorter individuals. Weight was included in the model to capture the hypothesis that evaluation scores would be affected by a preconceived notion of what Navy personnel should look like especially since the early 1980s when the US Navy began to cultivate a

culture of fitness and wellness. Additionally, a new variable was created by dividing the ratee's weight by his or her height. This variable was included to offset the fact that taller individuals will generally weigh more than shorter individuals. Dividing weight by height provided a better measure for individuals that are overweight by the Navy's cultural standards.

Education level was included to capture the effect of increased education. A similar variable to education level that was also included in the study was rater and ratee Standard Entry Test Score variables. Every enlisted individual entering the Navy must take a standard Armed Forces Qualification Test (AFQT) prior to their enlistment. The results of this test are used as a basis for detailing individuals into different job fields within the Navy and are often used as a measure of aptitude. The number of months an individual has been on board the ship was also included as an independent variable. This variable hoped to capture job experiences gained by being on board longer and as a measure for the rater and ratee familiarity. Some of the research has suggested that the more familiar the rater is with the ratee the higher the evaluation ratings will be. The number of times an individual has been disciplined at a non-judicial punishment (NJP) was included to capture an individual's discipline record. Due to the Navy's culture, NJP is considered a negative reflection of an individual's character and has a major impact on evaluation scores.

The rest of the variables were considered "similarity factors." These were factors that captured how much the rater's and ratee's demographic and personality measures were alike. It was hypothesized that the ratees that are more similar to the rater will receive higher evaluation scores than individuals that were less similar to the rater.

Studies have shown that individuals rate themselves higher than others do (Landy & Farr, 1980). It is reasonable to assume that most individuals that have advanced within an organization believe they possess good qualities. It is also reasonable to assume that if these raters see the same qualities in one of their ratees then they may rate that individual higher than individuals that do not possess the same like qualities. Examples of these similarity factors were: home of record, personality type, marriage status, number of children, and smoker/nonsmoker. Some previously mentioned variables were also looked at as similarity variables. These variables were: race, age, gender, education level, height, and weight.

Regression Model One

The first multiple regression model utilized the collected ratee characteristic data to quantify their effect on performance rating variation. In this model, variation in the individual's evaluation scores was decomposed into demographic and personality components. Initially, all the collected variables were inserted into the model, first as continuous and categorical and then as just categorical. From the best of these two models, all the non-significant variables, at a $p = 0.05$, were dropped. The effect of these components was examined to determine if they had a statistically significant effect on evaluation scores. This examination answered the first research question. Statistically significant nonperformance factors that are attributed to ratee characteristics and are determined to affect the accuracy of the rating process were then used in the third regression model.

Regression Model Two

In the second model, the existence of a rater/ratee similarity bias was specifically tested. In this model, variation in the individual's evaluation scores was decomposed into matched rater and ratee characteristic components. These matched components attempted to capture the raters' and ratees' similar demographic characteristics. This second model specifically answered the second research question and tested the hypothesis that individuals displaying like characteristics as the raters received evaluation scores that were statistically different from individuals that did not display similar characteristics. Since the Navy's evaluation process utilizes two raters, a method was required to capture all available combinations of rater and ratee matches. For example, the ratee may be similar to the first rater but different from the second rater in a certain category (race) while in a different category (gender), the ratee may be similar to the second rater and different from the first rater. To capture this, each category consisted of a series of dummy variables that reflected the following combinations of rater/ratee pairings:

1. First rater, second rater and ratee were all the same
2. First rater was the same as second rater but not the same as the ratee
3. First rater was the same as ratee but not the same as the second rater
4. Second rater was the same as ratee but not the same as the first rater
5. First rater, second rater and ratee were all dissimilar (this case was not applicable to the gender, the smoker/nonsmoker, children, and the married variables)

As an example for the Race variable, the pairings looked like the following:

R1 equal to one (1) if the race of first rater, second rater and ratee were all the same, otherwise R1 was equal to zero (0)

R2 equal to one (1) if first rater's race was the same as second rater but not the same as ratee, otherwise R2 was equal to zero (0)

R3 equal to one (1) if first rater's race was the same as ratee but not the same as second rater, otherwise R3 is equal to zero (0)

R4 equal to one (1) if second rater's race is the same as ratee but not the same as first rater, otherwise R4 is equal to zero (0)

R5 equal to one (1) if the rate, first rater and the second rater's race differ, otherwise R4 is equal to zero (0)

Again, only the factors that are statistically significant at a $p = 0.05$ were kept in the model and their effects on the rating scores were examined. The results of this model were then compared with the results of the first model. Specifically, variables that were found to be significant in both models were examined to see if the effects in the second model were greater or less than in the first model. As with the first model, statistically significant nonperformance factors that were attributed to matched rater and ratee characteristics were used in the third regression model.

Regression Model Three

Based on the results of the first two regression models, a final model was constructed. In this model, variation in the individual's evaluation scores was decomposed using the statistically significant components from both the first and second models. Only the factors that were statistically significant at a $p = 0.05$ were kept in the third model and their effects on the rating scores were once again examined. It was this model that answered the third research question. Ideally, differences in the actual performances of the ratees should be the primary factor that affects the variation in this performance rating cycle and nonperformance factors should have had little influence. Since this third model was designed using only nonperformance factors, the proportion of

total variation explained by these variables should be small. This proportion was indicated by the model's resultant coefficient of multiple determination or R-squared. Intuitively, the higher the R-squared becomes, the greater the chance that a rater bias exists.

Based on the R-squared level, an examination of the statistically significant variables was conducted to explain why they were significant in affecting the variation in performance score. In reference to Wherry's basic rating model, these variables represent the bias component in Wherry's basic equation. This third multiple regression model was then applied to the underlying theory of Wherry's rating equation and a determination was made as to whether or not the variation caused by these variables was attributed to rater biases and whether the study was successful in supporting Wherry's basic rating theory.

Findings

The findings presented in this section represent the results of the three regression models that were specifically constructed to test R. J. Wherry's fundamental rating equation. Before these findings are presented the sampling and data collection methods for this study are discussed followed by a description of the data sets used to construct the first and second models. The third model attempted to quantify all the collected nonperformance factors that affected the accuracy of this performance rating system. The results of this third model are then applied in the conclusion section of this study to Wherry's basic rating equation and his performance appraisal theory.

Data Collection and Sample Selection

The initial sample consisted of the entire population of the lowest three ranks of enlisted individuals on a single U.S. Navy aircraft carrier during one performance appraisal cycle. A total of 701 individuals were given an evaluation during this cycle and the results of these were collected directly from the evaluations. The primary data obtained from these evaluations were the ratees' performance scores and the identity of both raters. Other data collected included the department to which the ratees were assigned and the date they joined the crew of the carrier.

The next step in the selection process was the recording of the ratees' demographic data. This data was obtained directly from each ratee's service record. From the service records data on gender, education level, race, age, marriage status, number of children, home of record, date they joined the crew of the carrier, AFQT scores, and the number of occurrences of NJP were collected. Current height and weight data on the ratees were obtained from a physical readiness test that was conducted by the ship during the data collection period.

The ratee demographic data collection process was conducted over a three-month timeframe and was conducted immediately after the evaluation cycle. From the population of 701 individuals, four individuals were dropped from the study because their service records were not available from which to collect the demographic data. The most likely reason for their records not being available is that these individuals had departed from the ship before the completion of this portion of the data collection.

Concurrently with the collection of ratee demographic data, the collection of data on the ratees' and raters' personality types was accomplished. This collection process

required the distribution of a "Personality Profiler" to each ratee's supervisor and to each enlisted rater. Each supervisor was assigned the responsibility of providing the profiler to the ratees and then collecting and returning the profiler. As previously mentioned in this study, one additional question was added to the profiler requesting information on whether the individual smoked or not. Seven hundred and one profilers were sent out to capture this data and a total of 582 were returned. Of those that were returned, 37 profilers were given to individuals that did not receive an evaluation, 16 individuals returned the profiler but did not complete the profiler or omitted answering on one or more of the pages. Another 50 individuals returned the profiler but indicated an unwillingness to participate in the study by not filling out the profiler.

Collection of rater demographic data was conducted after completion of the collection of ratee demographic data. This collection process took an additional two months to complete. The data collected mirrored the demographic information taken from the ratees' service jackets with one exception. Information on NJPs was not available for raters. Additionally, the information on raters' education level and weight were not reliable due to the length of time between entering the data when they enlisted in the Navy and the time of the study.

A total of fourteen raters' data were not available for data collection due to their service records being unavailable. Collecting rater data after the collection of ratee data resulted in a three to five month time period where raters checked out of the command. Individuals that checked out took their records and therefore, their demographic data with them. An additional fifty-one ratees were dropped from the sample due to their rater's

service jackets not being available. Of the fourteen raters' data that were not available for data collection, three of them accounted for 32 of the 51 ratees that were dropped.

Of the original 701 individuals that received an evaluation, 423 had complete data sets and were kept in the sample for the models. Listed in Table 1 are the primary demographics from the sample of 423 individuals and from the 697 ratees that had service record data. This list demonstrates that the sample of 423 is an accurate representation of the larger sample. The biggest difference between the two groups is the percentage of smokers, 26 percent for the 697 ratees to 36 percent for the sample of 423. Since there was a specific question on the Personality Profiler that requested information on smoking, the sample of the 423 was considered the more accurate measure.

Table 1
Comparison of the Primary Demographics of the Sample of 423 Ratees and the Sample of 697 Ratees

Variable	423Avg/Percent	697 Avg/Percent
Eval Average	3.54	3.48
Age	21.84	21.89
Months On Board	17.89	17.83
AQFT	46.23	45.94
Height	68.05	68.08
Weight	162.45	163.28
Male	76%	77%
High School Degree	91%	90%
White	57%	56%
Black	28%	31%
NJP	20%	22%
Married	23%	22%
Kids	16%	15%
Smoke	36%	26%

The final step in the data collection process was to convert the raw demographic data on both the ratees and the raters into the data to be used in the models. This

conversion from raw data to formatted data resulted in the data sets that were specified in Chapter 3 of this study and were eventually used to construct the three regression models. Contained in Appendix A is a list of how each of the data was measured and labeled for the models.

Sample Demographics

Table 2 contains the sample demographics that were used in the study.

Table 2
Sample Demographics

	Category Variable	Number	Percent
Gender	Male	322	76%
	Female	101	24%
Education	High School Degree	383	91%
	Less than HS Degree	9	2%
	Some College	26	6%
	College Degree	5	1%
Race	Asian	23	5%
	White	243	57%
	Black	118	28%
	Other	22	5%
	Native American	17	4%
Others	NJP	84	20%
	Married	98	23%
	Kids	68	16%
	Smoke	159	36%

Continuous Variable	Mean	Std Dev	Min	Max
Eval Average	3.54	0.43	2.33	4.67
Age	21.84	2.45	17.75	35.75
Months On Board	17.89	9.98	2	55
AQFT	46.23	13.16	31	86
Height (Male)	69.21	2.98	60	79
(Female)	64.28	2.54	58	74
Weight (Male)	162.45	23.99	117	261
(Female)	143.85	22.33	86	238

As listed in Table 2, the 423 individuals that made up the sample had an average evaluation score of 3.54 out of a possible maximum score of 5.0. Other pertinent data

from this list are that over 75 percent of the ratees were males. Over 90 percent of the ratees had a high school diploma with only one percent attaining a college degree. Fifty-seven percent of the ratees were considered "White" and 28 percent were listed as "Black." The average age of this group was 21.84 with a minimum age of 17.75 and a maximum of 35.75.

Data not listed in Table 2 concerns the ratees' home of records and personality types. The ratees' home of records were as follows: 69 were from the Northeast, 59 were from the Southeast, 62 were from the Midwest, 17 were from the North Plain States, 71 were from the South Plain States, 14 were from the Northwest, 111 were from the Southwest, and 20 were considered not from the continental United States.

Table 3 list the ratees' personality types as they were determined to be from the following pairs: introverting (I) or extroverting (E) , sensing (S) or intuiting (N), thinking (T) or feeling (F), and organizing (Z) or adapting (A).

Table 3
Personality Types

Type	Number	Type	Number
ISFZ	27	ESEZ	29
ISFA	16	ESFA	16
ISTZ	29	ESTZ	39
ISTA	17	ESTA	25
INFZ	9	ENFZ	63
INFA	16	ENFA	54
INTZ	13	ENTZ	28
INTA	10	ENTA	32

Regression Model Number One

As the initial step in the data analysis, the first of three regression models was structured to identify statistically significant, nonperformance factors that influence the

variation in rating scores of an actual performance appraisal process. To accomplish this objective the ratee's average evaluation score was used as the dependant variable and regressed using the collected demographic data as independent variables. All the non-significant variables, at the $p = 0.05$ level, were then systematically dropped in a *Stepwise* manner. This process resulted in five significant variables remaining in the model with "White", "Months on Board, greater than one standard deviation below the average (MOBSD1)", "NJP", "Home of Record, Midwest (HORMW)" and "Smoke" as being significant. The result of this model is listed in Table 4.

Based on the literature on performance appraisal research the variables for gender, education, and age were all expected to have a measurable influence on evaluation scores and be statistically significant. To verify that these variables had no effect on evaluation scores, F-tests were performed. In each case gender, education, and age had no effect on the model and were rejected. Additionally, all continuous data were inserted into the model first as continuous variables and then as categorical data as specified in section 3 of this study. The categorical data did increase the model's R-squared value but only the variable for months on board was significant.

Table 4
Regression Model Number One: The Effect of Statistically Significant Ratee Demographic Data on Evaluation Scores

Number of Observations = 423 Prob > F = 0.00
R-squared = 0.15 Adj R-squared = 0.14

Variable	Coef.	t	P>|t|
MOBSD1	-0.30	-5.63	0.00
SMOKE	-0.08	-2.00	0.05
NJP	-0.13	-3.79	0.00
WHITE	0.15	3.86	0.00
HORMW	-0.12	-2.17	0.03

The R-squared of 0.15 indicates that this model captured approximately 15% of the variance in this appraisal process. The five variables that are contained in this model represent the statistically significant, nonperformance factors that influenced the variation in the rating scores of this evaluation cycle. An inspection of these nonperformance factors resulted in the following observations:

Months on Board, greater than one standard deviation below the mean (MOBSD1), had the greatest effect on the evaluation scores with a coefficient value of negative 0.30. This result was not surprising since the number of months on board for this category was eight months or less. Not only does the ratee perform better with experience but the rater also becomes more familiar with the ratee over time. Wherry himself acknowledged that "the longer the rater knows the ratee on the job, the greater the probability that the rating will be accurate" (Wherry & Bartlett, 1982). In acknowledgement of this, the Navy's evaluation system does not require an evaluation if an individual has been on board for less than 90 days.

The "White" variable had the next highest significance with a 0.15 effect on evaluation scores. This value was of interest as it indicates that white ratees tend to receive higher overall evaluation scores than nonwhites holding everything else constant. This seems to illustrate a possible bias towards white ratees or against nonwhite ratees.

The number of NJPs was anticipated to have a significant effect and this was confirmed in the model with a coefficient value of a negative 0.13. As previously stated, in the Navy's culture, receiving NJP is considered a reflection of an individual's character and generally has a major negative impact on evaluation scores.

The "Home of Record, Midwest (HORMW)" was significant having a negative 0.12 effect on evaluation scores. This value indicates that ratees from the Midwest received lower overall evaluation scores than ratees from all other regions. One possible explanation is that individuals from the Midwest also fell into one or more of the other categories in this model which had a negative effect on evaluation scores. There were 62 individuals in the study from the Midwest. Of these, 24 were nonwhite, 16 had gone to a NJP, 13 were on board for less than 8 months, and 31 were smokers. A total of 38 of the 62 fell into at least one of the four negative categories that were in the final model.

Smoking was the least significant variable of the first model with a negative 0.09 effect on evaluation scores. This indicates that ratees that smoke tend to receive lower overall evaluation scores than nonsmokers holding everything else constant. This seems to illustrate a possible bias against smokers.

Even though the intent of this model was to identify nonperformance factors that influence variation in performance evaluations, it is pertinent to also mention the factors that were found to be non-significant. Gender, age, marriage, having children, personality type, weight, and education level were all rejected as being significant in influencing the variance of this evaluation process. This indicates that for this evaluation, these variables do not show indications of rater bias.

Also of interest was the effect of converting continuous data into categorical data. In every case the variables that were converted improved the model's overall coefficient of determination. Additionally, groups outside one standard deviation from the average (both above and below) tended to receive lower evaluation scores than individuals that were within one standard deviation. Even though these variables were not significant at

$p = 0.05$, it does seem to indicate a possible bias towards individuals that are on the extremes of these variables.

Regression Model Number Two - Summary Statistics

The discussion of the statistical results for the second model is best began by examining the matches between the ratees and the raters. Table 5 lists these combinations of matches. The first column is a list of the variables that were matched. The second column represents the number for each variable where the ratees matched both the first and the second raters (R=R1=R2), the third column indicates the number of times the ratee matched the first rater but not the second (R=R1/R2). The third column indicates when the ratee matched the second rater but not the first (R=R2/R1). The fifth column shows the number of times where the ratee, the first rater, and the second rater differed (R/R1/R2). The last column indicates the number of times where there the first and second raters matched but they did not match the ratee for that variable (R1=R2/R). For the variables with only to possibilities (smoker or non-smoker), N/A was entered in column four.

Table 5
Regression Model Number Two: List of Matched Combinations Between Ratees and Raters

Variable	R=R1=R2	R=R1/R2	R=R2/R1	R/R1/R2	R1=R2/R
Gender	315	4	7	N/A	97
Smoke	154	92	91	N/A	86
Race	109	47	44	67	156
Age	67	78	49	80	149
AFQT	48	70	87	104	114
Type	13	21	22	273	94
Married	84	31	28	N/A	280
HoR	11	58	44	226	84
MoB	38	84	74	152	75
Children	0	25	0	216	182
Height	59	63	51	119	131

Regression Model Number Two

The second regression model was structured to identify statistically significant, nonperformance factors that influence the variation in rating scores by matching rater and ratee demographic characteristics. This model was designed to test the hypothesis that an individual displaying like characteristics as the raters will receive an evaluation score that is statistically different from an individual that does not display similar characteristics. As with all the models, the ratee's average evaluation score was used as the dependant variable. The independent variables for this model were sets of dummy variables that capture the possible matching combinations for each demographic category. Again, all non-significant variables at the $P = 0.05$ were dropped from the model. The results of the second model are presented in Table 6.

Table 6
Regression Model Number Two: Effect of Matched Variables on Evaluation Scores

Number of Observations = 423 Prob > F = 0.00
R-squared = 0.09 Adj R-squared = 0.08

Variable	Coef.	t	P>t
AgeRR2notR1	-0.14	-2.28	0.02
TypeRR1notR2	0.30	3.18	0.00
SmokeRR1notR2	0.11	2.60	0.01
RaceRR2notR1	0.26	3.82	0.00
KidsDiffer	0.14	3.39	0.00

The R-squared of 0.09 indicates that this model captured less than 10% of the variance in this appraisal process. The five variables that are contained in this model represent the statistically significant, matched nonperformance factors that influenced this variation in rating scores. Four of the five variables for this model had a positive effect on

evaluation scores. A more detailed inspection of these matched nonperformance factors results in the following observations:

Two of the variables, "Age" and "Race", indicate that ratees that matched the second raters received evaluation scores that differed from ratees that did not match. The variable that matched the ratee's race with the race of the second rater but not the first rater (RaceRR2notR1) had the greatest positive effect and was the most significant with a 0.26 effect on evaluation scores. This is interesting in that it indicates that ratees with the same race as the second rater receive evaluation scores that are higher than ratees that are not, holding all other factors constant.

Concerning the age matched variable (AgeRR2notR1), it seems to indicate that ratees within the same age category as the second rater receive evaluation scores that are lower than ratees that are not, holding all other factors constant. There is not an obvious reason for this relationship.

Two of the variables, "Type" and "Smoke", indicate that ratees that matched the first raters received evaluation scores that differed from ratees that did not. The variable that matched the ratee's personality type with the personality type of the first rater but not the second rater (TypeRR1notR2) had a positive effect on evaluation scores and was significant with a coefficient value of 0.30. This is interesting in that it indicates that ratees with the same type as the first rater receive evaluation scores that are higher than ratees that are not, holding all other factors constant.

The same effect is also reflected in the matched smoker or nonsmoker category where ratees that match the first rater (SmokeRR1notR2) receive higher evaluation scores than those that do not match.

The only variable that indicates that it is beneficial to differ from either the first or second rater is in the area of having the same number of children. When the ratee did or did not have the same number of children as the first and second raters (KidsDiffer) they received higher evaluation scores than ratees that matched the raters. Again, there is not an obvious explanation for this result, however, there were only 25 cases where the ratee matched the first rater in this category and there were no incidences where the ratee matched the second rater. The infrequency of matches may explain these results if ratees that matched the first raters also happened to receive lower evaluation scores on average. Of the 25 matches between ratee and the raters, 21 of those ratees also had a NJP, were on board less than 8 months, smoked, were from the Midwest or were not White. The average performance score for those 21 individuals was 3.32 which was well below the sample average of 3.54.

Even though the R-squared value was relatively low, the model seems to support the hypothesis that individuals displaying like characteristics as the raters will receive evaluations scores that are statistically different from individuals that do not display similar characteristics. The matched variables for "Type", "Smoke", and "Race" all support this hypothesis and indicate that matching the raters will result in higher evaluation scores. However, the matched "Age" and "Kids" variables do not support the hypothesis and seems to indicate the opposite effect might be true.

Regression Model Number Three

The third regression model used the statistically significant, nonperformance factors that influence the variation in rating scores identified in models one and two. This model was designed to test the hypothesis that these previously identified

nonperformance factors will have a statistically significant effect on the evaluation scores and that a portion of this variation can be attributed to rater bias. Again, the ratee's average evaluation score was used as the dependant variable. The results of this third model are presented in Table 7 after all the non-significant variables were dropped. This model's R-squared increased to 0.20 (up from 0.15 in model one and 0.09 for model two.) This model has approximately 20 percent of the variance in this appraisal process. The eight variables that are contained in this model represent the statistically significant, nonperformance factors that influenced this variation in rating scores.

Table 7
Regression Model Number Three: Effect of Models Number One and Two Significant Variables on Data on Evaluation Scores

Number of Observations = 423 Prob > F = 0.00
R-squared = 0.20 Adj R-squared = 0.18

Variable	Coef.	t	P > t
TypeRR1notR2	0.19	2.20	0.03
White	0.14	3.63	0.00
RaceRR2notR1	0.18	2.78	0.01
KidsDiffer	0.14	3.69	0.00
Smoke	-0.09	-2.17	0.03
MOBSD1	-0.28	-5.36	0.00
NJP	-0.13	-3.82	0.00
HORMW	-0.11	-1.97	0.05

All five of the variables contained in the first model are included in this model. Months on board, greater than one standard deviation below the mean has the greatest effect on the evaluation scores with a coefficient value of -0.28. This result reflects that ratees with little to no experience on the job tend to receive lower evaluations than their peers that have more experience and time on the job.

The number of NJPs was significant with a -0.13 effect on evaluation scores. The "Home of Record, Midwest" variable is significant in this final model with a -0.11 effect

on evaluation scores. The "Smoke" variable had a -0.09 effect on evaluation scores. These negative values indicate that ratees that smoke, went to NJP or are from the Midwest tend to receive lower overall evaluation scores holding everything else constant.

Three of the five variables contained in the second model are included in this model. The variable that matched whether an individual smoked or not and the variable the matched age groups were dropped in the final model. The variables matching race, and personality type were retained, as well as, the variable that indicated the ratees and raters differed in having children.

The variable that matched the ratee's personality type with the personality type of the first rater had a positive 0.19 effect on evaluation scores. This result indicates that ratees with the same type as the first rater receive evaluation scores that are higher than ratees that are not, holding all other factors constant. Just like the second model, ratees that differed from the first and second raters in the number of children they had tended to receive higher evaluation scores.

The "White" variable was significance with a positive 0.14 effect on evaluation scores. This coupled with the variable that matched the ratee's and rater's races hold the greatest possibility of reflecting a rater bias. These two factors indicate that white ratees receive higher evaluation scores and if the ratee's race matches that of the second rater they will tend to receive higher scores as well holding all other factors constant.

Overall, this model supports the hypothesis that previously identified nonperformance factors do have a statistically significant effect on the evaluation scores and that a portion of this variation can be attributed to rater bias.

Conclusion

Personal judgments by raters are central to the performance appraisal process; however, these judgments often result in an inaccurate measure of a ratee's performance. As a result, a substantial amount of research has been conducted in an attempt to improve the accuracy of performance appraisals. In 1952, R. J. Wherry developed his rating process theory. Key to his theoretical work was a fundamental rating equation, which stated that a rating score was equal to the actual performance of the ratee plus an observation and recall bias component plus random error. Wherry argued that this equation on the rating process provided a method of measuring rater bias. However, little research has been conducted to test the validity of or to improve on Wherry's rating process. As such, this study utilized the basic foundation of his theory and regression analysis to quantify the nonperformance factors that affected the accuracy of an actual rating process. These nonperformance factors are representative of the bias component in Wherry's equation. This section of the paper will draw conclusions from the results of the three regression models that were used to identify this bias component. These conclusions will be followed by a discussion of the limitations of the study and a section on recommendations for future research.

The first model attempted to isolate the effects of ratee demographic characteristics on the variation in performance rating scores. Five of the demographic variables were found to be statistically significant, and they captured approximately 15% of the variance of this appraisal process. The five variables represented individuals that either were on board for less than eight months, smoked, were white, committed a military offense, or were from the Midwest.

The second model examined the effect of similar rater and ratee demographics to see the extent to which these matches influenced variation in performance scores. The R-squared of 0.09 indicated that this model captured less than 10% of the variance in this appraisal process. The model specifically tested the hypothesis that individuals displaying like characteristics as the raters will receive evaluation scores that are statistically different from individuals who do not display similar characteristics. The matched variables for "Type", "Smoke", and "Race" all supported this hypothesis and indicated that ratees matching the raters in these three areas resulted in higher evaluation scores.

Based on the results of the first two regression models, a final model was constructed. This final regression model decomposed variation in the ratees' evaluation scores using statistically significant components from both the first and second regression models. The final model's R-squared value of 0.20 indicated that these previously identified nonperformance factors had an influence on the variation in evaluation scores. Of the eight variables that are contained in the third and final model there were four variables that strongly indicated the existence of rater bias. Ratees that were either white, had personality types that matched the first raters, or were of the same race as the second raters all received higher evaluation scores than ratees that were not. Additionally, ratees that smoked received lower evaluation scores.

The finding that white ratees receive 0.14 higher performance scores on average than nonwhites was consistent with the findings of Landy and Farr's report on performance rating. They cited six of seven studies that found white ratees receiving higher evaluation scores than black ratees. The one exception found no difference in evaluation scores (Landy & Farr, 1980). These studies, supported by the findings of this

study, clearly indicate a possible bias towards white ratees. However additional study is required to determine causality; for example, it is possible that white ratees enter the Navy better qualified to succeed based on quality of their education, positions of previous leadership, or social expectations. Nonetheless, this finding represents a disparity in evaluation scores between the races that needs to be addressed through rater education. It is also possible that nonwhite ratees may require specific training once they enter the Navy to account for previous social differences that benefit white ratees.

The finding that indicated that ratees who matched the race of the second rater received 0.18 higher evaluation scores was also consistent with the findings of Landy and Farr. They reported that ratees tended to receive higher ratings from raters of their own race (Landy & Farr, 1980). This finding, coupled with the finding that showed a rating bias towards the variable that matched the ratee's personality type, supports the hypothesis that there is a specific rater bias towards individuals that display similar characteristics as the rater. Turban and Jones (1988) found that rater and ratee characteristic similarity appeared to be positively related to supervisor evaluations of subordinates. The results of this research support their findings that rater and ratee similarity positively influences evaluation scores. In their study they stressed that "more research was needed to understand the mechanisms by which similarity influences evaluation" (Turban & Jones, 1988).

The "Smoke" variable indicated a possible negative bias towards smokers. As with the previous variables, additional research is needed to determine causality. Lower scores may not necessarily be a bias against smokers but a reflection of lower scores given to smokers because they are away from the work space more often during smoke

breaks. Related to this issue was the second model finding that revealed that ratees that matched the first rater as to smoking or not smoking preference received 0.11 higher scores. This seemed to further support the rater/ratee similarity hypothesis. Additional research is needed to determine if ratee smokers who had a first rater who also smoked received higher average evaluations than nonsmokers. However, this finding seemed to add support to the argument that there is a bias against smokers.

Overall, this model supported the hypothesis that previously identified nonperformance factors have a statistically significant effect on the evaluation scores and that a portion of the variation in these scores can be attributed to rater bias. The result of this third multiple regression model did support the underlying theory of Wherry's rating equation in that rater biases can be identified and measured empirically.

Even though the intent of this model was to identify nonperformance factors that influence variation in performance evaluations, it is worth mentioning the factors that were found to be non-significant. Gender, age, marriage, having children, personality type, weight, height, and education level were all rejected as being significant in influencing the variance of this evaluation process. This study indicated that rater bias does not exist in these areas, at least not for this sample. This finding is especially noteworthy concerning age and gender where a good deal of research on the effect of demographic nonperformance data on evaluations has centered (Hartel, Douthitt, Hartel, & Douthitt, 1999; Landy & Farr, 1980; Bigoness, 1976; Hamner, Kim, Baird & Bigoness, 1974; Pulakos, White, Oppler, & Borman, 1989; Nevill, Stephenson & Philbrick, 1983). For this evaluation cycle, the nonperformance factors for gender and age did not influence evaluation scores and therefore are not a source of rater bias.

Additionally, converting continuous data into categorical data should seriously be considered when empirically studying performance appraisals. In every case the variables that were converted to categorical data, using one standard deviation to define the categories, improved the model's overall fit. This method seemed to capture biases that occur for or against individuals that fall outside the "normal" range that might be masked when using continuous data. Even though these variables were not significant at the $p = 0.05$ level, the model's results did seem to indicate a possible bias towards individuals that are outside the normal range of these variables. This finding would be consistent with the study performed by Hamilos and Pitz (1977) which found that subjects discriminated between old and new data more effectively when the data presented was towards the extreme maximum or minimum values.

Policy Implications

The existence of rater bias in a performance appraisal process is clearly undesirable and methods to mitigate it must be developed. Most likely the majority of the bias in this appraisal system is unconscious and can be reduced significantly through education. A strategy to educate Navy raters on the two primary biases, race and familiarity, found in this study must be seriously considered. For example, a pilot study could be conducted along with the education to determine the effectiveness of the training. Regression models should be estimated before and after the training on a specific unit for sequential evaluation cycles. If the training proves effective in reducing rater biases uncovered during the first cycle then the training should be provided Navy wide.

However, if a bias towards white ratees still exists after the education intervention then a specific study designed to examine the differences in social backgrounds of white and nonwhite ratees should be conducted. This study would look at factors that are deemed important in order to receive high evaluations, and these would be compared to social and professional skills of white and nonwhite ratees. If this study resulted in finding a difference between white and nonwhite ratees, then ratee training and education for nonwhite ratees should also be considered. This training and education should be designed to mitigate those differences.

Even though more research is still needed to determine the factors that may have produced these biases, corrective measures could still be implemented now to mitigate these potential rater biases. This strategy is supported by the literature; for example, Wherry stated that if the rater makes a conscious effort to be objective after becoming aware of a biasing influence, then the rater should be able to reduce the influence of that bias (Wherry & Bartlett, 1982).

Limitations

The first limitation to this study was that it captured the results of a single performance appraisal cycle on board a single ship. Although the nonperformance factors that influenced variation in this cycle were identified and measured, generalizations to other appraisal cycles are somewhat limited; additional quantitative and qualitative study is required to establish causation. Therefore, recommendations to improve this performance rating system are restricted.

Another limitation is that the findings of this study cannot be applied directly to other organizations. Variables that influenced the variance in this study may not have any

influence in other organizations, which limits the relevance of this study to other organizations. However, even if the findings are not directly transferable, the methods used to capture these findings can be applied to other performance appraisal systems.

Additionally, this study is a snapshot study that only covers one performance rating cycle of one small segment of a very large organization. The findings that exist for this cycle may not exist in future appraisal cycles or in other segments of this organization. This limits the ability to make broad inferences or generalized statements based on the findings of this study. Before such statements could be made, further study of the Navy organization's performance rating process, conducted over a number of rating cycles, would be required.

Future Research

This research has exposed at least five areas where future research is required. The first area that needs additional study is whether white ratees enter an organization better "qualified" to succeed. Factors that lead to promotion need to be identified. Once they have been identified, these factors would then have to be compared to the qualifications that entry level ratees possess. If white ratees have an initial advantage, then measures would need to be developed to change the promotion factors or to provide training to the ratees that lack the proper qualifications to succeed in the organization.

The second area of additional study concerns rater and ratee similarity and how similarity influences the evaluation process. Previous research has indicated that ratees rate themselves above average when compared to fellow workers (Bartol, Durham, & Poon, 2001). A possible extension of this research is that individuals that perceive similar qualities in their subordinates may also rate those ratees above those who did not possess

similar qualities. Further research is required to test this hypothesis. Additionally, as Turban and Jones (1988) pointed out, similarity between ratee and rater may produce a working environment where the ratee is more confident and has more insight into what is needed to receive a better evaluation. It is possible that it is this insight and not rater bias that causes higher evaluation scores.

Additional study is needed on the effect that smoking has on evaluation scores. The finding that smokers are given lower evaluation scores than nonsmokers needs to be replicated in other studies. If it is indeed true, then research will be required to determine why. A more extensive look at whether raters that smoke give higher evaluation scores to ratee smokers than ratees that do not smoke and if raters that smoke give smokers higher evaluation scores than raters that do not smoke is needed as well.

Another important caveat is that the research conducted in this study needs to be expanded to include a greater portion of the Navy. To obtain a clearer picture of the Navy's performance appraisal process these models need to be re-estimated on other ships and units over multiple evaluation cycles. If the biases found in this study are consistently replicated throughout the Navy then methods to mitigate the biases would need to be developed. Finally, this study used only the most basic principles of Wherry's performance rating equation. The results of this study do support these principles but further empirical study of performance appraisals will need to be done to continue testing and substantiating Wherry's theories. Additional research must be concentrated on the factors that influence the variation in performance in order to identify and measure their effect. Once they have been identified, then adequate controls can be developed and empirically tested for effectiveness.

References

Bartol, K. M., Durham, C. C., & Poon, J. (2001). Influence of performance evaluation rating segmentation on motivation and fairness perceptions. *Journal of Applied Psychology, 86*, 1106-1119.

Bernardin, H. J., & Villanova, P. (1986). Performance appraisal. In E. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 43-62). Boston: Heath/Lexington.

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

Borman, W. C. (1977). Consistency of rating accuracy and rater errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20*, 238-252.

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

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86,* 965-973.

Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel selection. *Annual Review of Psychology, 48,*299-338.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*, 130-135.

DeNisi, A. S., & Peters, L. H. (1996). Organization of information in memory and the performance appraisal process: Evidence from the field. *Journal of Applied Psychology, 81*, 717-737.

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

Hamilos, C. A., & Pitz, G. F. (1977). The encoding and recognition of probabilistic information in a decision task. *Organizational Behavior and Human Performance, 20,*184-202.

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

Hartel, E. J., Douthitt, S. S., Hartel, G. & Douthitt, S. Y. (1999). Equally qualified but unequally perceived: Openness to perceived dissimilarity as a predictor of race and sex discrimination in performance judgments. *Human Resources Development Quarterly, 10*, 79-94.

Hogan, E. A. (1983). *The formation of Supervisor's initial expectations and subsequent expectation effects on performance appraisals: A longitudinal field study.* Dissertation. University of California, Berkeley.

Keeping, L. M., & Levy, P. E. (2000). Performance appraisal reactions: Measurement, modeling, and method bias. *Journal of Applied Psychology, 85,* 708-723.

Kirsch, M. P., & Zalesny, M. D. (1986, August). *Effects of rater-ratee similarity on performance ratings*. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

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

Lefkowitz, J. (2000). The role of interpersonal affective regard in supervisory performance ratings: A literature review and proposed causal model. *Journal of Occupational and Organizational Psychology, 73,* 167-189.

Madden, J. M., & Bourdon, R. D. (1964). Effects of variation in rating scale format on judgment. *Journal of Applied Psychology*, 48, 147-151.

Matens, J. (1999). Conducting effective performance reviews. *Modern Casting* (Vol. 89, No. 6, pp. 54-56.

Mobley, W. H. (1982). Supervisor and employee race and sex performance appraisals: A field study of adverse impact and generalization. *Academy of Management Journal, 26,* 598-606.

Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. *Journal of Applied Psychology, 71*, 39-44.

Nevill, D. D., Stephenson, B. B., & Philbrick, J. H. (1983). Gender effects on performance evaluation. *Journal of Psychology, 15,* 165-169.

Oberg, W. (1999). *Make performance appraisal relevant.* http://www.unep.org/restrict/pas/pasta.htm (Retrieved December 10, 2002)

Pulakos, E. D., & Wexley, K. N. (1983). The relationship among perceptual similarity, sex, and performance ratings in manager-subordinate dyads. *Academy of Management Journal, 26,* 129-139.

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

Schmitt, N., & Lappin, M. (1980). Race and sex determinants of the mean and variance in performance ratings. *Journal of Applied Psychology, 65,* 428-435.

Schoorman, F. D. (1988). Escalation bias in performance appraisals: An unintended consequence of superior participation in hiring decisions. *Journal of Applied Psychology, 73,* 58-62.

Smith, P. C., & Kendall, L.M. (1962). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47, No. 2,* 149-155.

Strauss, J. P., Barrick, M. R., & Connerly, M. L. (2001). An investigation of personality similarity effects (relational and perceived) on peer and supervisor ratings and the role of familiarity and liking. *Journal of Occupational and Organizational Psychology, 74,* 637-657.

Turban, D. B. & Jones, A. P. (1988). Supervisor-Subordinate similarity: Types, effects, and mechanisms. *Journal of Applied Psychology, 73,* 228-234.

Wandelken, D. J., & Inn, A. (1981). Nonperformance influences on performance evaluations: A laboratory phenomenon? *Journal of Applied Psychology, 66,* 149-158.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35,* 521-551.

Williams, K. J., DeNisi, A. S., Blencoe, A. G., & Cafferty, T. P. (1985). The role of appraisal purpose: Effects of purpose on information acquisition and utilization. *Organizational Behavior and Human Decision Processes, 35,* 314-339.

Appendix A

List of Variables

List of Variables

Variable	Category Breakdown (Label for Study)
Race	White (White), Black or African American (Black), Asian (Asian), American Indian/Alaska Native (AmerIndian), Other or Unknown (Other)
Gender	Dummy variable with Male = (1) and Female = (0) (Gender)
Height	Continuous data measured in inches (Hgt)
Weight	Continuous data measured in ounces (Wgt)
Age	Continuous data measured in years and fraction of year (Age)
Education level	Less than High School Degree (LessThanHS), High School Degree (HSDegree), Some college (SomeCollege), College Degree (ColDegree)
Number of months assigned to the ship	Continuous data measured in months (MOB)
Standard entry test scores	Continuous data from service records (AFQT)
Discipline record	Dummy variable with record of nonjudicial punishment = (1), otherwise = (0) (NJP)
Personality type	Sixteen possible combination from four categories as determined by the personality profiler (ISFZ, ISFA, ISTZ, ISTA, INFZ, INFA, INTZ, INTA, ESFZ, ESFA, ESTZ, ESTA, ENFZ, ENFA, ENTZ, ENTA)
Married/not married	Dummy variable with married = (1), otherwise = (0) (Married
Number of children	Dummy variable with children = (1), otherwise = (0) (Kids)
Smoke/Nonsmoker	Dummy variable with smoker = (1), otherwise =(0) (Smoke)

Home of record

North East	New Jersey, Maryland, Virginia, New York, Vermont, Rhode Island, Maine, New Hampshire, Connecticut, Pennsylvania, Delaware, District of Columbia, or Massachusetts (HORNE)
South East	West Virginia, Georgia, Florida, South Carolina, North Carolina, Tennessee, Mississippi, Louisiana, or Kentucky (HORSE)
Midwest	Ohio, Indiana, Minnesota, Michigan, or Illinois (HORMW)
North Plains	Wisconsin, Idaho, North Dakota, Iowa, South Dakota, or Montana (HORNP)
South Plains	Arkansas, Kansas, Nebraska, Oklahoma, Colorado, New Mexico, or Texas (HORSP)
North West	Washington, Utah, or Oregon (HORNW)
South West	Arizona, California, or Nevada (HORSW)
All Others	Hawaii, Alaska, Guam, Virgin Islands, etc. (NONCON)

Note: For the continuous data that was converted to categorical data, the continuous data was broken down into the following categories; less than one standard deviation below the mean, within one standard deviation below the mean, within one standard deviation above the mean, or more than one standard deviation above the mean. The labels for these variables; Age, Months Assigned to the Ship, Standard Entry Test Scores, Height, and Weight, were as follows:

AgeSD1 = less than one standard deviation below the mean ratee age
AgeSD2 = within one standard deviation below the mean ratee age
AgeSD3 = within one standard deviation above the mean ratee age
AgeSD4 = greater than one standard deviation above the mean age.