

DOES COMFORT MATTER? THE ROLE OF RATER DISCOMFORT ON
PERFORMANCE RATINGS

A dissertation submitted to the faculty of
The California School of Professional Psychology
In partial fulfillment of the requirements for the degree of
Doctor of Philosophy
At Alliant International University, Los Angeles, California

By
Mina Azizi

May 2014

UMI Number: 3624600

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3624600

Published by ProQuest LLC (2014). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

© 2014

Mina Azizi

All Rights Reserved

ALLIANT INTERNATIONAL UNIVERSITY Los Angeles

The dissertation of Mina Azizi, directed and approved by the candidate's Committee, has been accepted by the Faculty of the California School of Professional Psychology in partial fulfillment of the requirement for the Degree of

DOCTOR OF PHILOSOPHY

DATE

Dissertation Committee:

Nurcan Ensari, Ph.D., Chairperson

Denise Lopez, Ph.D., Member

Jonathan Troper, Ph.D., Member

ROLE OF RATER DISCOMFORT ON PERFORMANCE	iv
TABLE OF CONTENTS	
LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF APPENDICES	xi
VITA	xii
ABSTRACT	xiii
CHAPTER I	1
Statement of the Problem	1
Causes of Rating Errors	3
Trait Variables	6
Conscientiousness and agreeableness	7
Self-efficacy	7
Raters' Discomfort	9
The Level of Performance	10
The Present Study and Research Questions	10
Definitions of Key Variables	12
CHAPTER II	14

ROLE OF RATER DISCOMFORT ON PERFORMANCE	v
Literature Review	14
Performance Management and Appraisal	14
Use of performance appraisal in organizations	15
Practical benefits of performance appraisal	19
Organizational and human resources decision making	19
Feedback and performance improvement	22
Performance appraisal methods	23
Alternatives to performance management and appraisal	29
Critique of performance appraisal	34
Rater Bias in Performance Appraisals	37
Unintentional biases in ratings	38
Intentional biases in ratings	42
Predictors of Performance	45
Conscientiousness	49
Agreeableness	51
Rater discomfort	53
Self-efficacy	54
The level of performance	57
Summary and Conclusion	59
CHAPTER III	61
The Present Study	61

ROLE OF RATER DISCOMFORT ON PERFORMANCE	vi
Importance of the Study	61
Purpose of the Study and Hypotheses	64
Hypothesis 1	65
Hypothesis 2	67
Hypothesis 3	70
Hypothesis 4	72
Hypothesis 5	74
Hypothesis 6	76
CHAPTER IV	79
Methods	79
Participants	79
Protection of human participants	81
Power analysis	82
Design	85
Procedure	85
Instruments	89
Measure of conscientiousness and agreeableness: IPIP NEO	89
Measure of new general self-efficacy: NGSE	91
Measure of rating discomfort: PADS	91
Measure of performance	92

ROLE OF RATER DISCOMFORT ON PERFORMANCE	vii
CHAPTER V	94
Results	94
The Results of the Pilot Study	94
Descriptive Statistics	95
Test of Assumptions	98
Independence of cases	98
Normality	98
Test of sub-group differences	100
Linearity and homoscedasticity	100
Tests of Hypotheses	100
Hypothesis 1	100
Hypothesis 2	101
Hypothesis 3	106
Hypothesis 4	109
Hypothesis 5	109
Hypothesis 6	114
CHAPTER VI	118
Discussion	118
Interpretation of Findings	118
Hypothesis 1	118

ROLE OF RATER DISCOMFORT ON PERFORMANCE	viii
Hypothesis 2	119
Hypothesis 3	120
Hypothesis 4	121
Hypothesis 5	122
Hypothesis 6	123
Relationship of Current Findings to Previous Research	124
Limitations and Considerations	126
Implications and Future Research	130
REFERENCES	134

LIST OF TABLES

1. Procedure overview	88
2. Descriptive Statistics	96
3. Correlations among experimental measures	97
4. Kolmogorov-Smirnov tests of normality	99
5. Mean scores for low performance and high performance conditions	103
6. Mean levels of performance rating across performance conditions and rater discomfort groups	107
7. Results of Bootstrapping Analysis for Hypothesis 3	110
8. Results of Bootstrapping Analysis for Hypothesis 4	112
9. Results of Bootstrapping Analysis for Hypothesis 5	115
10. Mean levels of performance rating across conscientiousness and rater discomfort groups	117

LIST OF FIGURES

1. The predictive relationship between the level of performance and performance ratings	66
2. The research model for Hypothesis 2	68
3. Moderational effect of the level of performance about performance on the relationship between rater discomfort and the level of performance rating	69
4. Research model for Hypothesis 3	71
5. Research model for Hypothesis 4	73
6. Research model for Hypothesis 5	75
7. Research model for Hypothesis 6	77
8. The moderational role of conscientiousness on the relationship between rater discomfort and performance ratings	78
9. Bar graph displaying means for low and high performance conditions	102
10. Mean performance ratings across levels of rater discomfort and information about performance	105
11. Rater discomfort's insignificant mediation on the relationship between conscientiousness and performance rating	108
12. Rater discomfort's insignificant mediation on the relationship between agreeableness and performance rating	111
13. Rater discomfort mediating the relationship between self-efficacy and performance rating.	113
14. Mean performance ratings across levels of rater discomfort and conscientiousness	116

LIST OF APPENDICES

A. Sample Qualtrics Survey Panel Invitation	155
B. Written Invitation Sent to Potential Participants	156
C. Employee Performance Vignettes	157
D. Employee Performance Rating Form	159
E. Demographics	160
F. Pilot Study	161
G. Histograms	164

VITA

Mina Azizi was born in Philadelphia, Pennsylvania. She received a B.S. in Business Management and a B.S. in Psychology from Pennsylvania State University in 2006.

ABSTRACT

Performance ratings are widely used in organizations to inform decisions on pay, promotion, training, and other organizational functions. The current study investigates how rater discomfort affects the assignment of performance ratings as well as potential mediating and moderating effects of raters' individual differences, namely conscientiousness, agreeableness, and self-efficacy. Participants were given one of two vignettes describing a fictional employee's performance and were asked to assign a rating using a six-item measure of employee performance. Participants also provided responses to measures of conscientiousness, agreeableness, self-efficacy, and rater discomfort. The level of performance developed for either vignette was shown to determine performance ratings. Additionally, raters with lower levels of rater discomfort were found to assign more extreme ratings while raters with high discomfort tended to provide ratings closer to the center of the scale. No other significant mediating or moderating effects were found on the relationship between the level of performance and performance rating. These findings confirm previous research indicating that rater discomfort can have a profound effect on performance ratings and may introduce measurement error due to unintentional rater leniency or other unwanted extraneous factors. Practical implications include the importance of organizational awareness of individual rater's differing comfort level, and experience in producing fair employee performance ratings. As such, it is recommended that organizations identify raters who have high discomfort and are in need of additional training and coaching. Future research should focus on expanding on the role of rater discomfort using information gathered from real employees rather than paper people.

CHAPTER I

An overview of the study is presented in this chapter, which begins with a statement of the problem and then gives a brief overview of the relationship among the key variables. This chapter also presents the research questions the study aims to address, as well as a list of definitions of the key variables.

Statement of the Problem

Performance appraisal is the formal procedure that an organization uses to assess job performance of employees. Employee performance appraisal, whereby a manager evaluates and judges the work performance of other employees, is one of the most common management practices utilized in organizations. Over 90 percent of large organizations employ some performance appraisal system (Bernthal, Sumlin, Davis, & Rogers, 1997), while over 75 percent of state employment systems require annual performance appraisals (Murphy & Cleveland, 1991). Performance appraisals are a human resource management tool, used widely in today's organizations to determine organizational and individual effectiveness and job promotion. This assessment data can also be used for administrative decisions, and employee development and feedback. Administrative decisions are important because many jobs specify performance data as a reason for pay raises or termination of employment. Employee development and feedback is equally important because employees can benefit from knowing how they are performing. The increasing importance of feedback has led to more companies integrating performance feedback into their human resource practices. Most companies conduct performance appraisals annually, but now they seem to be going beyond this by

designing a more comprehensive performance management system (Muchinsky, 2012).

In addition to annual appraisals, these systems often include employee and supervisor goal setting, as well as coaching and feedback sessions between employee and supervisor (Spector, 2003).

Despite the increased effort to make performance management systems robust and their ever growing popularity, often times it is difficult for organizations to evaluate whether their performance appraisal system is accomplishing the desired outcomes. Although practitioners have implemented changes to rating instruments, evaluation criteria and the appraisal procedure in an effort to improve the accuracy and perceived fairness of the process, the practice is still under scrutiny. Many are still dissatisfied with the system; employees sometimes view it as inaccurate and unfair. Many times this inaccuracy materializes as rating elevation, where ratings are largely restricted to the positive end of the continuum on a rating scale.

Barrett (1966) considered the problem common to "virtually every rating program. When a program is initiated, more than half the people are given ratings above average and the proportion of high-rated people grows until only the obvious misfits fail to make the top grades" (p. 23).

The negative perception of performance appraisals largely stems from the reality that many managers are inconsistent in applying objective criteria to performance appraisals, resulting in unreliable and sometimes deliberately distorted evaluations (Folger, Konovsky, & Cropanzano, 1992). Rating bias in performance appraisals can have a number of negative consequences. Distortions in performance ratings undermine

the integrity of an appraisal system, and erode work motivation, commitment and loyalty. When performance ratings are linked to pay or bonuses, then falsely-elevated ratings can exhaust the available funds designated for merit increases within organizations, and may alter the company pay-for-performance structure in the future. For example, performance standards will be higher or more difficult to achieve in organizations in which ratings are routinely inflated. Furthermore, employee training needs cannot be recognized or identified when ratings are falsely-inflated. Moreover, the blemishes of the organization as a whole can be covered up by inflated ratings and the opportunity for organization diagnosis and development may be lost. For example, if all the members of a department are performing at a high standard, according to inflated performance ratings, then the decision-makers of the company may be misled by the ratings into ignoring potential problems within the department. For these reasons, amongst others, it is critical to the success of a performance appraisal system that the process be fair and valid.

Causes of Rating Errors

All of the criticism and concern over the accuracy of performance appraisals has been a catalyst for researchers and practitioners to search for the underlying causes of inaccuracy and rating errors. In past studies of performance appraisals, researchers have focused on accuracy. Rating errors are understood to be the result of the rating stimuli that do not trigger reliable and valid responses (Cronbach, 1955). From a cognitive perspective, rating errors are conceptualized to be the result of the limitations of human cognition (DeNisi & Peters, 1996), such as memory accessibility (Murphy & Balzer, 1986), cognitive style (Cardy & Kehoe, 1984; Härtel, 1993), and affect (Cardy &

Dobbins, 1986). This is because human judgment, which is a large part of performance ratings, tends to be imperfect. When leaders or supervisors make performance ratings, they are likely to exhibit rating biases and rating errors. These biases and errors can be seen in the pattern of ratings, both within the rating forms for individuals and across ratings forms for different employees. These within and across-form patterns are called halo and distributional errors, respectively.

Halo error occurs when a rater gives an individual the same rating across all dimensions, despite differences in performance across dimensions (Balzer & Sulsky, 1992; Cooper, 1981). For example, a police officer might be outstanding in completing many arrests (high quantity) but do a poor job in paperwork. A supervisor might rate this officer high on all dimensions, even though the uniform high rating is not deserved or accurate. This error can also occur if an employee performs poorly in one area, then the employee is marked as poor in all areas, even though his or her performance may be satisfactory on other performance dimensions. The concern with halo error is explaining the cognitive processes that would lead a rater to exhibit halo error. Several researchers have theorized that raters rely on a general impression of the employee when making dimension ratings (Lance, LaPointe, & Stewart, 1994; Nathan & Lord, 1983). According to this view, salient pieces of information are used to form an impression of an employee. The impression then forms the basis for performance ratings. This suggests that raters may not be suited to administer ratings on performance dimensions, but instead are suited to an overall performance score. This is not to say that an employee cannot perform at a uniformly high or low level on all standards. Halo patterns might accurately indicate that

dimensions of actual performance are related. This possibility has led to considerable discussion in the industrial-organizational psychology literature about the meaning of halo (e.g., Balzer & Sulsky, 1992; Murphy & Jako, 1989; Murphy, Jako, & Anhalt, 1993; Pulakos, Schmitt, & Ostroff, 1986; Solomonson & Lance, 1997). Part of this discussion concerns how to separate the error from true halo. True halo occurs when an employee actually performs at a similar level on all dimensions.

Distributional errors occur when a rater tends to rate every employee similarly. Distributional errors include leniency, severity, and central tendency errors. Leniency errors occur when the rater rates every employee at the desirable end of the scale while severity errors occur when the rater rates every employee at the undesired end of the scale (Bass, 1956; Hauenstein, 1992; McIntyre, Smith, & Hasset, 1984). Central tendency errors occur when a rater rates every employee in the middle of the performance scale (Murphy & Balzer, 1986).

While many raters involuntarily commit rating errors due to poor scale design or poor training, raters sometimes do intentionally distort ratings so as to achieve some specific goals (Murphy & Cleveland, 1991). Willful acts of rating distortion tend to occur at the time the rating is rendered and not when performance information is observed, encoded, or recalled (Kane, 1994) and are motivated by a variety of reasons (Villanova & Bernardin, 1989). For example, leniency in ratings may be motivated by raters wanting to achieve a more harmonious work group and an avoidance of discomfort when rating the employee. One study found that in the U.S., ratings are commonly skewed more as one moves up the organizational hierarchy (Bretz, Milkovich, & Read,

1992). A common reason found for rating inflation is to obtain rewards for subordinates like promotions and salary increases (Lawler, 1976; Longenecker, Sims, & Gioia, 1987).

Other researchers have presented ideas that outline the complexity and the potentially biased nature of appraisal ratings. Whisler (1958) suggested that the utility of performance appraisals as an accurate measure of employee performance is limited due to the hesitancy of raters to be completely honest for fear of negative repercussions. Harris (1994) argued that in order for raters to rate accurately, they need to be motivated to do so. Predictors of rater motivations in Harris' framework include such things as situational factors (e.g., being accountable to a supervisor), negative consequences (e.g., damage to subordinate-supervisor relationship), and rewards (e.g., likelihood of pay increases or promotions).

Trait Variables

In addition to voluntary distortion, performance rating errors may result from limitations in raters' skills and cognitive capacities (DeNisi & Peters, 1996; Feldman, 1981; Landy & Farr, 1980), as well as the rater's individual or personality characteristics, such as self-efficacy, agreeableness, and conscientiousness.

Personality is defined as the combination of stable physical and mental characteristics responsible for a person's identity (Kreitner & Kinicki, 2007). Personality traits can be defined as habitual patterns of behavior, thought, and emotion (Kassin, 2005). There seems to be an infinite number of potential traits used to describe personality, but the statistical technique of factor analysis has demonstrated that particular clusters of traits reliably correlate together (Eysenck, 1991). Many

psychologists believe that five factors adequately describe human personality (McCrae & Costa, 1987) and that personality traits are relatively stable over time, differ across individuals, and influence behavior (Kreitner & Kinicki, 2007).

Conscientiousness and agreeableness. Conscientiousness and agreeableness are two personality traits that have been shown to be predictors of biased performance ratings (Bernardin, Cooke, & Villanova, 2000). Conscientiousness and agreeableness are two dimensions of the Five Factor Model (FFM) of personality. Individuals high in conscientiousness are characterized as hard working, thorough individuals with high standards (McCrae & Costa, 1987). Individuals high in agreeableness are characterized as sympathetic, cooperative, and desire social approval (McCrae & Costa, 1987). Past research has predicted, and found, that raters high in conscientiousness are less prone to rating elevation (Bernardin et al., 2000). Researchers also found that raters high in agreeableness tend to produce elevated ratings (Bernardin et al., 2000). It should be noted that no significant relationship has been found between the remaining three FFM personality traits (openness to experience, neuroticism and extraversion) and performance ratings.

Self-efficacy. The self-efficacy of the individual executing the performance appraisal is another trait variable that has been found to influence performance ratings. Raters' self-efficacy, within the context of performance management, refers to the raters' belief that he or she may orchestrate performance in the course of fulfilling their role obligation (Bernardin & Villanova, 2005). In other words, self-efficacy, as it pertains to

performance appraisal, is the degree to which a rater believes he or she can competently perform the performance appraisal duties of his or her job.

Self-efficacious raters believe themselves capable of executing socially demanding behaviors that have important consequences for their relationships with ratees (Bernardin & Villanova, 2005). In the performance appraisal process, raters may be concerned in cases when they need to administer or give unfavorable or lower ratings to an employee. More self-efficacious raters may better organize information concerning employee information and performance to provide more compelling justifications for rating according to a standard (Bernardin & Villanova, 2005). So, high self-efficacy raters may be more determined in applying these standards when evaluating employee performance than less self-efficacious raters (Bernardin & Villanova, 2005). However, the most important behavioral outcome of rater self-efficacy is reduction of rating inflation (Bernardin & Orban, 1990; Bernardin & Villanova, 2005). Raters who are self-efficacious tend to be stricter in their evaluations of others. Simply put, rater self-efficacy provides the necessary courage to evaluate others accurately. This courage has been found to stem from a rater's management of the different aspects of the process (Bernardin & Villanova, 2005). First, more efficacious raters rely on higher quality information and are more confident in their ability to provide compelling justifications for their evaluations. Second, they are more resolute in applying performance standards. Third, they are better able to provide useful performance improvement information to ratees. Finally, they are more capable of performing those social behaviors related to the successful resolution of conflict (Bernardin & Villanova, 2005).

Raters' Discomfort

Along with trait characteristics like personality variables and self-efficacy, state variables, which are temporary ways of interacting with self and others, can also affect performance ratings. One such state variable related to performance evaluations is rater discomfort, which occurs when an employee experiences uneasiness, heightened anxiety or withdrawal when charged with rating a subordinate for the purposes of a performance appraisal (Villanova & Bernardin, 1989).

The concept of rater discomfort was born from job compatibility theory. Job compatibility theory refers to the extent to which employees maintain preferences for job characteristics that are consistent with the actual demands of the job (Villanova, Bernardin, Dahmus, & Sims, 1993). According to the job compatibility framework, employees whose preferences are at odds with their job characteristics tend to report greater discomfort in performing job activities and manifest behaviors indicative of less job involvement and higher withdrawal and avoidance (Villanova et al., 1993; Spence & Keeping, 2009). Conducting performance appraisals is not a job demand that is typically consistent with employee preferences and job characteristics. Researchers have found that raters frequently report discomfort with several facets of the performance appraisal process, including the monitoring of employee performance, evaluating performance, and providing performance feedback (Murphy & Cleveland 1991). Villanova et al., (1993) noted that raters who show high levels of appraisal discomfort are more likely to provide inflated ratings and are less likely to distinguish among employees. In other words, raters

may be motivated to assign uniformly high ratings in order to avoid discomfort associated with making difficult judgments of others' performance.

The Level of Performance

In addition to the roles of trait and state variables, the level of performance is an important factor that can have an effect on performance ratings. Previous research has assumed that raters are likely to have different levels of rating inflation for ratees with different performance levels (Wong & Kwong, 2007). However, while this assumption has been formally tested, results are not well established. One of the major purposes of performance evaluation is to discriminate among employees who perform at various levels. This kind of differentiation is relevant to major personnel decisions (Murphy & Cleveland, 1995), and rating validity in terms of differential elevation (Cronbach, 1955). Thus, there is a gap in the literature regarding how various levels of performance are evaluated differently across different rater conditions. Furthermore, ratees' performance level is an important contextual factor shaping rating behaviors (Gaugler & Rudolph, 1992; Wexley, Sanders, & Yukel, 1973; Wong & Kwong, 2007). For example, research on contrast effects has shown that the performance of a target ratee is often contrasted with the performance of the preceding candidate (Maurer & Alexander, 1991; Wexley et al., 1973).

The Present Study and Research Questions

Due to the popularity of performance evaluations in organizations as well as the propensity for distorting them, it is important to further examine the factors that may influence managers when doing performance reviews. Furthermore, performance ratings

have many implications in the workplace, and can provide valuable performance information to a number of critical human resource activities, such as the allocation of rewards, like merit pay and promotions. Performance appraisals can also provide feedback on the development and assessment of training needs, selection predictors, and performance documentation (Cleveland, Murphy, & Williams, 1989). When it comes to the intangible, appraisal systems can hold potential for enhancing the effectiveness of human resource decisions and for satisfying employee needs for performance feedback (Ilgen, Fisher, & Taylor, 1979).

From past research, we know that there is a strong relationship between performance ratings and both conscientiousness and agreeableness (Bernardin et al., 2000). However, mediating and moderating relationships between these variables have yet to be explored. The current study aims to expand the current literature by examining (a) the mediating role of rater discomfort on the relationship between conscientiousness and performance ratings, (b) the mediating role of rater discomfort on the relationship between agreeableness and performance ratings, (c) the mediating role of rater discomfort on the relationship between self-efficacy and performance ratings, and (d) the moderating role of the level of performance on the relationship between rater discomfort and performance ratings.

The present study contributes to academic literature and applied settings in many ways: First, it examines the relationships among several well-researched variables in the performance appraisal literature. The mediating and moderating relationships among the key variables have not been examined in a single study before. Second, it can assist

organizations to administer specific appraisal training in an attempt to overcome rating bias, and develop employee training, such as self-efficacy training for raters, to eliminate this bias. Last, it can lead to increased trust and faith in company appraisal processes which in turn may increase employee motivation.

The research questions are as follows:

1. Does the level of performance affect the performance rating?
2. Does the level of performance moderate the relationship between rater discomfort and the level of performance rating?
3. Does rater discomfort mediate the relationship between conscientiousness and the level of performance rating?
4. Does rater discomfort mediate the relationship between agreeableness and the level of performance rating?
5. Does rater discomfort mediate the relationship between self-efficacy and the level of performance rating?
6. Does conscientiousness moderate the relationship between rater discomfort and the level of performance rating?

Definitions of Key Variables

Performance Rating – An evaluative rating given to an employee by his/her supervisor, representing the supervisor's assessment of the employee's work performance.

Conscientiousness – the quality of acting according to one's conscience. It includes the components of self-discipline, carefulness, thoroughness, organization,

deliberation, and need for achievement (McCrae & John, 1992).

Agreeableness – a tendency to be accommodating and pleasant in social situations. This trait is based on trust, straightforwardness, altruism, compliance, modesty and tender mindedness (McCrae & John, 1992).

Rater Discomfort – the level of anxiety and uneasiness one experiences when tasked with evaluating an employee in a performance appraisal (Villanova et al., 1993).

Self-Efficacy – in the context of performance management, refers to the raters' beliefs that they may orchestrate performance in the course of fulfilling their role obligation as it pertains to performance management (Bernardin & Villanova, 2005).

The level of performance – explicit information about how well an employee is performing.

Performance Appraisal Experience – number of years of appraisal experience someone has had.

CHAPTER II

Literature Review

This chapter will provide an overview of the existing literature surrounding the topic of rater bias in performance ratings. The review will begin with a background on the performance management process and the use of performance appraisals in organizations. This will be followed by a review of the critique within performance appraisal research and an overview of alternatives to performance appraisals systems in the current organizational climate. The chapter will conclude with a review of factors that contribute to rating biases.

Performance Management and Appraisal

Performance management is a strategic, organization-wide plan to formally assess employee performance, in order to improve business operations, reduce inefficiencies, and reduce costs associated with operations and human capital. Performance management includes a cycle of events, which consist of a systematic process of planning and setting goals on the organizational and individual levels, continually monitoring and systematically rating performance, giving feedback and rewarding or penalizing good or poor performance (McNamara, 2005). The first phase of performance management involves setting goals. This includes establishing the elements and standards of performance appraisal plans. Effective Human Resource (HR) systems strive to make these elements measurable, understandable, verifiable, equitable and achievable (United States Office of Personnel Management, 2011). The next phase in the performance management cycle, performance appraisal, involves consistently monitoring and

measuring performance. Within the context of formal performance appraisal requirements, rating means evaluating employee or group performance against the elements and standards in an employee's performance plan, and assigning a summary rating (United States Office of Personnel Management, 2011). This rating is typically assigned based on procedures included in the organization's appraisal program. When managers assign ratings to employees, it is typically based on work performed during an entire appraisal period. In most organizations, the appraisal period equals one year, while other organizations may use a continual performance management model with shorter and more frequent appraisal periods (Muchinsky, 2012). The next phase in the performance management cycle involves providing ongoing feedback to employees and workgroups on their progress toward reaching goals, with the intent to improve future performance (Billikopf, 2006). Although employees vary in their desire for improvement, generally they at least want to know how well they are performing (Kubo & Saka, 2002). Once this feedback is shared with employees, a successful performance management initiative involves the implementation of performance improvements and returns to phase one to revise the key performance indicators to be measured in the subsequent performance management cycle.

Use of performance appraisal in organizations. Performance appraisal systems have been a common element in the workforce since 1914, when Lord and Taylor Co. instituted a formal performance evaluation system, in which they started rating their employees annually against pre-established performance objectives (Markle, 2000). Depending on time and industry practices, performance appraisal systems have been

called performance reviews, annual reviews, performance appraisals, merit ratings, performance ratings, and employee ratings (Markle, 2000). Historically, most appraisals were designed for managers to assess employee commitment to the organization, their contribution to projects, and skills like communication and teamwork (Milkovich & Wigdor, 1991).

Performance appraisals are used throughout the world, in all sectors of business including private sector, and for and non-profit organizations of various sizes (Tziner, Murphy & Cleveland, 2001). Murphy and Cleveland (1991) reported several studies indicating that 74-89 percent of the surveyed organizations had a formal appraisal system. In 1995, William Mercer Inc. surveyed 218 companies, and determined that almost all management and technical and knowledge workers received annual performance evaluations (Markle, 2000). A British study revealed that 82 percent of the participating organizations operated some formal performance appraisal process (Long, 1986). In addition, for nearly 50 years, the United States federal government has operated with some performance appraisal procedures whose purposes have been to strengthen the link between pay and performance (Milkovich & Wigdor, 1991). Not a great deal has changed as appraisals continue to assess much of the same dimensions as they always have. One major thing that has indeed changed is the nature of jobs in our society (Tziner et al., 2001). Traditionally, performance was based on a known output quality, volume, dollar value or even responsiveness (Neely, Richards, Mills, Platts, & Bourne, 1997). Much of the American workforce was employed in manufacturing and blue-collar jobs where there was a tangible and obvious product to measure employee performance

(Neely et al., 1997). Today, many jobs are in the service sector or do not produce a tangible product (e.g., customer service, business consultant, computer engineer). Even with this major change, performance ratings continue to have many implications in the workplace.

Among researchers, it is agreed upon that the purpose of performance appraisal has been well established (Murphy & Cleveland, 1995). This involves allocating individual outcomes such as merit raises or promotions (individual oriented) and identifying needs in human resource planning, training and development, and organization of work (collective oriented). Sources of evaluation typically include supervisors or managers, peers, subordinates and a 360-degree technique with only employees or both employees and their managers being evaluated (Beer, 1978; Murphy & Cleveland, 1995).

Coens and Jenkins (2002) identified five elements common to almost all performance appraisal systems: (a) the performance, behaviors or traits of individuals are rated or judged by someone else; (b) these ratings are scheduled, usually annually or quarterly, as opposed to being tied to completion of particular tasks or projects; (c) such ratings are not applied to selected individuals, but rather are systematically undertaken with all employees of a particular department; (d) the process is either strictly mandatory or tied to some reward system; (e) information is recorded and kept in the employee's file by the employer. Primarily, ratings provide valuable human resource information for organizations to allocate rewards, like merit pay and promotions (Schraeder & Jordan, 2011). Ratings also provide feedback on the development and assessment of training

needs, selection predictors, and performance documentation (Cleveland et al., 1989).

When it comes to the intangible, appraisal systems can hold potential for enhancing the effectiveness of human resource decisions and for satisfying employee needs for performance feedback (Ilgen et al., 1979). Because of its critical importance in enhancing both employee and organizational performance, performance appraisal is considered a central human resource activity in organizations (Atwater, Wang, Smither & Fleenor, 2009; Levy & Williams, 2004; Murphy & Cleveland, 1995). Generally, both employees and organizations benefit from performance appraisal. On one hand, employees receive opportunities for feedback, development and rewards; while organizations benefit from being able to monitor individual employee performance and link performance to strategic business goals (Claus & Briscoe, 2009).

Although the importance of performance appraisals within organizations has long been recognized, in more recent years, researchers and practitioners have found performance appraisals to be a controversial and polarizing issue. There is abundant evidence in the psychological literature that the contexts in which performance ratings are obtained and used do not yield objective ratings (Tziner, Murphy & Cleveland, 2005). This is to say that researchers must carefully consider the impact of contextual factors such as rater personality, organizational norms, beliefs and opinions (Tziner et al., 2005). Before considering the controversy that surrounds objectivity and performance appraisal research and practice, it is important to understand benefits and established practices regarding performance appraisal.

Practical benefits of performance appraisal. Performance appraisal is a vehicle to validate and refine organizational actions, such as selection and training (Billikopf, 2006). Performance ratings are primarily used to measure the performance of employees and make strategic business decisions (Cleveland et al., 1989). Organizations and their employees, both raters and ratees, benefit from using performance appraisals. The most obvious benefit for organizations is that they fill a need for performance measurement. In fact, appraisals aid organizations by supporting goal achievement in four main areas: (a) administrative purposes (e.g., decisions about promotions, remuneration, or dismissal), (b) employee development, (c) assessment of employee potential, and (d) research purposes (e.g., use as criterion; Drenth, 1998; Murphy & Cleveland, 1991). In addition, performance appraisal allows organizations to set up a process where judgments are formalized and structured. This is especially important because research has concluded that there is a basic human tendency to make judgments about those one is working with (Dulewicz, 1989). In the absence of a structured appraisal system, people will tend to judge the work performance of others informally and arbitrarily. This human inclination to judge can create serious motivational, ethical, and legal problems in the workplace. A structured appraisal system gives an opportunity for these judgments to be fair and lawful.

Organizational and human resources decision making. From the perspective of many organizations and human resource functions, performance appraisals are investments that yield many different positive results (Campbell & Pritchard, 1976). First and foremost, a performance appraisal system informs managers and organizations for

the purpose of promotion and compensation. This is based on the idea of accountability. When employees are aware that their organization is mindful of their performance and that they will be rewarded with merit increases, promotions or other opportunities; they are motivated to work harder (Ryan & Deci, 2000). In addition, if a reward system is put in place, then morale improves when employees receive these positive rewards for their work (Podsakoff, MacKenzie, Paine, & Bachrach, 2000). Professional employee development is considered another benefit of performance appraisals (Pollack & Pollack, 1996). Having a systematic procedure of documenting employee performance allows organizations with an opportunity to address performance problems and to analyze the strengths and weaknesses of employees (Billikopf, 2006). In addition, positive performance or high performing individuals can be identified (Schraeder, Becton, & Portis, 2007). The organization is then in a position to effectively utilize the skills of all their employees. Through improving training and promoting high performers, employees can perform their jobs at the highest level and be in a better position to do their job (Spinks, Wells, & Meche, 1999). This is based on the thinking that a well-developed staff is more likely to be proactive, productive and resourceful, all of which helps give an organization a competitive advantage. Performance appraisals also help managers to frame the validity of their selection procedures. An effective performance appraisal system can assist an organization in achieving its goals and objectives and identify training needs (Spinks et al., 1999). A performance appraisal system can also bring about more enhanced communication and improved employee morale. In addition to the aforementioned benefits, performance appraisal can also identify gaps in performance

and causes of performance deficiencies, hence resulting in improving performance (Kramar, McGraw & Schuler, 1997). As mentioned, performance appraisal systems can provide valuable information for managers on many fronts, and employees have access to feedback about their performance and effectiveness (Campbell & Pritchard, 1976). Lastly, organizations can use performance appraisal as a criterion measure to complete a strategic departmental benchmarking initiative, for example. All of these benefits ultimately contribute to the so-called bottom line.

As mentioned earlier, organizations of all kinds, including small service firms, nonprofit organizations, government institutions and public and private companies typically participate in performance appraisal. Early empirical studies found links between firm performance and individual HR policies, such as compensation (Gerhart & Milkovich, 1990; Gomez-Mejia & Balkin, 1992) and employee selection (Terpstra & Rozell, 1993). This supported emerging attention on the importance of HR decisions in understanding organizational performance. More recent research has established a link between a broader array of HR policies and organizational performance (Delery & Doty, 1996; Huselid, 1995; Huselid, Jackson, & Schuler, 1997). The effect sizes in these studies have been substantial in practical terms. For example, in the three studies just cited, measures of accounting profits or cash flow were about 20 percent higher on average in organizations having HR practices that were one standard deviation above the mean on dimensions such as HR effectiveness and that included what has become known as high performance work practices: pay for performance, participation in decisions, investment in training, and so forth (Delery & Doty, 1996; Huselid, 1995; Huselid et al., 1997).

Feedback and performance improvement. Employee feedback sessions provide an opportunity for a manager or HR professional to give employees information regarding their performance to identify areas of strength and weakness, and to facilitate performance improvement. In most organizations, feedback sessions are conducted privately and face-to-face between employee and manager (Geake, Oliver & Farrell, 1998). Feedback may be qualitative or quantitative or a combination of the two (Billikopf, 2006). Some researchers feel feedback is particularly useful when workers have an achievement objective (Billikopf, 2006). Performance improves substantially (11% to 27%) in a number of settings when workers were given specific goals to achieve, and received performance feedback (Latham & Locke, 2007). In one case, managers observed that truck drivers seldom loaded their trucks more than 58% to 63% of capacity. After goals were set to load trucks to 94% of capacity, truckers achieved an average of 80% capacity within the first month and were frequently surpassing 90% within the first three months. As a result, the company saved an excess of \$250,000 within a nine-month period (Latham & Locke, 2007). In this example, management communicated an expectation of performance in order to facilitate performance improvement.

It is important for researchers and practitioners to acknowledge the interconnectedness of each part of the performance management process, from goal-setting and rating to feedback and development. When researching a specific segment of the process, such as ratings, it must be noted that it is inevitable for the ratings process to be affected by the norms and practices of the segments surrounding it. Researchers have also tied effective feedback to important organizational outcomes, such as job satisfaction,

employee learning, and motivation (Hackman & Oldham, 1976; Mignerey, Rubin, & Gorden, 1995; Morrison, 1993; Murphy & Cleveland, 1995; Wanberg & Kammeyer-Mueller, 2000).

Performance appraisal methods. The usefulness of a performance management system hinges on the success and accuracy of the performance appraisal process. In order to accurately assess the key performance indicators that have been identified as critical to the organization's operations and objectives, a rating method must be chosen that can appropriately and effectively measure said performance indicators. When designing or selecting a rating instrument, important considerations include (a) how the rating method will affect how ratings are calculated and used, (b) the ease with which managers can learn to use the instrument, and (c) how business goals and improvement processes can be tied into the performance rating results (Billikopf, 2006; United States Office of Personnel Management, 2011). The link between organizational effectiveness and performance can be assessed at multiple levels (individual, group, plant, business unit, and firm). It is generally believed that different levels of analysis are useful (Becker & Huselid, 1998; Delery & Shaw, 2001). In this vein, studies at the plant or facility level (e.g., Arthur, 1994; Ichniowski, Levine, Olson, & Strauss, 2000; MacDuffie, 1995; Youndt, Snell, Dean, & Lepak, 1996) have also found important relationships between HR practices and performance. However, it is the prospect of a link between HR practices and organizational performance that has become of greatest interest to researchers and managers. In order to derive information from performance appraisal that

can drive organizational and HR decisions, organizations must choose instruments that provide meaningful information to both employees and management (Billikopf, 2006).

In order to assess employees in a systematic and efficient fashion, organizations typically use a specific process. Most performance appraisals used today are based on a rating scale of some kind. Not only are rating scales the most common performance appraisal method, but they are sometimes used in conjunction with other methods to yield a more robust performance appraisal. Ratings scale methodology requires an employer to develop an in-depth grading system, similar to the way students in school are assessed (Billikopf, 2006). This scale is then used to evaluate employee performance within a variety of areas, such as technical skill set, teamwork and communication skills (Billikopf, 2006). Typically, there is a minimum grade an employee must receive in order for the performance appraisal to be considered a success. It is not uncommon for managers to use a performance improvement plan for employees who do not meet this cutoff criterion. A rating scale method is viewed by some management theorists as an egalitarian way of measuring individual performance (Billikopf, 2006).

Several types of appraisal data gathering exist. The most popular used in organizations include objective production and judgmental evaluation.

The objective production method consists of direct, but limited, measures such as sales figures, production numbers, and electronic performance monitoring of data entry workers (Muchinsky, 2006). Depending on the job and its duties, the measures used to appraise performance would vary. Accidents and absenteeism can also serve as useful indicators of job performance (Muchinsky, 2006). This type of appraisal data gathering

suffers from criterion contamination and criterion deficiency (Muchinsky, 2006). In other words, because the variability in performance can be due to factors outside of the employee's control and because the quantity of production does not necessarily indicate the quality of the products, the objective production method is usually incomplete in gathering appraisal data (Muchinsky, 2006).

Judgmental evaluation involves individuals evaluating the performance of others. Many organizations use evaluation forms with a rating scale that includes pre-determined anchors. Raters make their judgment by noting a check or circle for the most appropriate rating that reflects the performance. Anchor-based appraisals include rating factors with a numerical scale (e.g., 1 to 7), or an adjective-descriptive scale (e.g., superior, good, below average) (Billikopf, 2006). The use of predetermined anchors is often combined with a narrative, in which a qualitative assessment of employee performance is documented. This qualitative judgment uses steps that are similar to steps taken when using a critical incident method. The critical incident technique will be outlined in the next paragraph.

One popular approach to performance appraisal, which falls under the category of the judgmental evaluation method, is the critical incident technique where supervisors are told to recall instances where employees reacted particularly well or poorly. Many times, this technique is used immediately after a critical incident has occurred. To be effective and accurate, critical incidents need to be documented as they occur and are still fresh in the supervisor's mind (Carroll & Schneier, 1982). Examples of negative critical incidents include significant errors made while performing job tasks and illegitimate absenteeism.

Examples of positive critical incidents include the obtainment of better-than-expected sales volume and performing a complicated job task in a particularly effective way.

The critical incident technique consists of a set of procedures for collecting direct observations of human behavior in such a way as to facilitate their potential usefulness in solving practical problems and developing broad psychological principles (Flanagan, 1954). It should be noted that critical incident technique is a procedure for gathering certain important facts concerning behavior in defined situations (Flanagan, 1954).

Because this technique works best in critical situations, it is not an appropriate method for an all-encompassing appraisal system, although it can be a helpful tool to include as part of a larger performance appraisal system in certain work environments that have many critical incidents (ex. Manufacturing). This technique is also used in organizational development as a research technique for identification of organizational problems because it deemphasizes the inclusion of general opinions about management and working procedures, and focuses on specific incidents instead (Flanagan, 1954).

The strength of this process is in the concreteness of the incidents documented by managers (Carroll & Schneier, 1982). Other advantages include its low-cost and ability to provide rich qualitative information. The pitfalls of this process include the likelihood that supervisors may emphasize negative worker behavior, especially if managers are not aware of the importance of recalling both positive and negative incidents; the possibility that some workers may be quite steady and not produce any particularly good or poor behavior for long periods of time; and since critical incidents rely on memory, incidents may be imprecise and remembered wrong by users or may even go unreported.

Additionally, in situations where a critical incident is not appraised immediately, this method has a built-in bias towards incidents that happened recently, since they are easier to recall. Lastly, respondents may not be accustomed to or willing to take the time to verbally explain or write a complete story when describing a critical incident. In addition to disadvantages for the direct users of the technique, there is also a disadvantage for the HR and strategic business leaders formulating the performance appraisal process of the organization. When critical incident technique is used alone, HR leaders may have difficulty translating critical incident reports into improved day-to-day performance (Carroll & Schneier, 1982). A judgmental method that includes a rating scale can potentially make for more standardized evaluations than the critical incidents approach and is less time consuming for managers (Billikopf, 2006). At the same time, a benefit for HR leaders is that the critical incident approach can be used to generate data and ideas to develop more complex rating scales (Carroll & Schneier, 1982). In reality, a combination of approaches is often necessary to end up with a useful performance appraisal (Billikopf, 2006).

Other types of performance appraisal include informal one-to-one review discussions, observation on the job, job-related skill tests, assessment centers, and psychometric tests (United States Office of Personnel Management, 2011). It should be noted that none of these methods are mutually exclusive. All of these performance assessment methods can be used in conjunction with others in the list, depending on the situation and organizational policy.

Behaviorally-anchored rating scale (BARS) is an appraisal method that aims to combine the benefits of narratives, critical incidents and quantified ratings by anchoring a quantified scale with specific narrative examples of good, moderate and poor performance (Schwab, Heneman, & DeCotiis, 1975). Behavior-based rating formats are generally superior to other formats in fostering performance improvement; when used with performance feedback, they tend to facilitate clarification of work roles for employees and the reduction of role ambiguity and conflict (Tziner & Falbe, 1990). At the same time, the same stream of research has found forced-choice scales to be better able to minimize deliberate rating inflation, making them preferable for administrative purposes such as promotion, merit pay, and employment termination (Tziner & Falbe, 1990).

Another appraisal procedure that incorporates multiple assessment approaches is 360-degree feedback, also referred to as multi-source feedback (Atkins & Wood, 2002). With this method, the rater interviews an employee, as well as the employee's supervisor, peers, self, and any direct reports (Fleenor & Prince, 1997). Multi-source feedback has increased in popularity and its popularity been facilitated by the increased use of web-based surveys on the Internet (Atkins & Wood, 2002). About 28 percent of HR professionals surveyed said their companies used 360-degree feedback as part of the review process and of those that did not, about 74% said there is no plan to implement such a program in the next year (Freedman, 2006). Multi-source appraisal techniques allow an appraiser to gain a more complete performance profile of the employee. In addition to assessing employee job performance and skill sets, an appraiser can receive

in-depth feedback on the employee's behavior, character, and leadership skills, which can be useful information to manage and help develop the employee (Bracken & Paul, 1993). It should be noted that multi-source feedback is an appraisal method that can employ the rating scale method and other types of appraisal. Multi-source feedback have attracted a good amount of research attention in the last decade and the majority of 360-degree feedback studies focus on issues such as self-other agreement and the impact of 360-degree feedback on behavioral change (Atwater, Waldman, & Brett, 2002; London & Smither, 1995).

Most individuals concerned with performance measurement depend on judgmental indices of one type or another (Landy & Farr, 1980). Between 1950 and 1955, 81% of the published studies in the *Journal of Applied Psychology* and *Personnel Psychology* used ratings as research criteria (Guion & Gottier, 1965). Blum and Naylor (1968) sampled articles from the *Journal of Applied Psychology* during 1960 to 1965 and found that of those using criterion measurement, 46% measured performance via judgmental indices. Landy, Farr, Saal, and Freytag (1976) reported that 89% of 196 police departments in major metropolitan areas used supervisory ratings as the primary form of performance measurement. Finally, Landy and Trumbo (1980) reported that a literature review of validation studies in the *Journal of Applied Psychology* between 1965 and 1975 revealed that ratings were used as the primary criterion in 72% of the cases.

Alternatives to performance management and appraisal. Alternative methods of achieving the objectives of traditional performance management and appraisal have been developed and used in organizations. One such alternative to performance ratings is

total quality management (TQM). TQM focuses on teams instead of only the individuals within organizations (Engholm, 1998). TQM has an overriding focus on the organization's customers, process and culture. The methods for implementing this approach came from the teachings of quality leaders such as Phillip B. Crosby, W. Edwards Deming, Armand Feigenbaum, Kaoru Ishikawa, and Joseph Juran (Engholm, 1998). At its core, TQM is a management approach to long-term success through customer satisfaction. In a TQM effort, all members of an organization participate in improving processes, products, services and the culture in which they work (Deming, 1986). A core concept in implementing TQM is Deming's 14 points, a set of management practices to help companies increase their quality and productivity. Some of these practices include on-the-job training and eliminate numerical quotas for the workforce and numerical goals for management (Deming, 1986). This style of management has been instituted in over 3,000 corporations and 40 government institutions in the United States (Milakovich & Wigdor, 1991). In TQM, the entire organization is considered a system of interlocking processes where the institution rather than the employees are considered the object of management. Despite the conversion to TQM, most of those using it persist in managing performance through individual employee ratings, a practice antithetical to TQM (Booz, Allen, & Hamilton, 1982; Usilaner & Leitch, 1989). According to TQM, problems do not originate with employees, but from a lack of understanding of the work processes. TQM is most easy to apply to a production or supply chain business, so it would not be practical for a law firm, for example, to adopt TQM.

One of the difficulties managers and bureaucrats have in following Deming's advice is that performance evaluations are so entrenched in the administrative mindset that it is inconceivable to eliminate them (Law, 2007). Deming is a famous critic of performance appraisal and when asked what an organization should do in place of performance appraisals, Deming is reported to have replied: "If your performance evaluation system does more harm than good, just quit doing it. You don't have to have an alternative to make an improvement" (Markle, 2000, p. 6). Deming's answer reflects a growing audience of researchers and practitioners who are seeking alternatives to the current HR practices of carrying out performance appraisals. Many management professionals suggest that the solution is to create better appraisal programs or alternatively to consider the information from appraisals within a wider context, along with other sources of information (Brinkerhoff & Kanter, 1980). As a result, organizations frequently revamp their performance appraisal systems. One study revealed that over seventy percent of companies surveyed had either changed their system in the last two years, or intended to do so in the future, and reported that companies often restructure the performance appraisal systems two or three times a decade (Markle, 2000). Some organizations that have given up individual performance appraisals have replaced them with alternatives that evaluate the performance of work groups or teams instead (Law, 2007). This still represents a top-down, judgmental management style that is so similar to individual performance appraisal it should not be considered an alternative (Lawler, 1994). Similarly, some organizations have reverted to the old practice of gift-giving in place of their evaluation-based merit systems (Law, 2007).

According to Kohn (1999), reward systems are still a form of external control and fail to motivate individuals. In 1957, before introducing the concepts of Theory X and Theory Y, Douglas McGregor published a critique of performance appraisals and offered an alternative approach. This essentially involved a paradigm shift, in that McGregor called for self-appraisal by an employee, rather than external evaluation by a manager (McGregor, 1957). This approach starkly contrasts with the predominant school of psychology thought of that era (e.g., behaviorism). McGregor's Theory Y provides a different approach to external control management, offering a substantially different type of relationship between managers and employees. The open communication and trusting relationship removes the need for formal performance evaluations. Other writers have proposed management approaches similar to McGregor. Markle (2000) used the term catalytic coaching to describe a management style which has at its core a partnership between employee and manager characterized by open, two-way communication and a shared vision of one another as capable, motivated individuals. Similarly, Peters and Waterman (1982) called for the empowerment of employees by expanding their opportunities for self-direction and self-control. Scholtes' total quality leadership, which is rooted in the ideas of Deming, called for "a fundamentally different view of the relationship between employees and the organization" (Joiner & Scholtes, 1988, p. 4). Scholtes replaces the notion of management with that of leadership, where from the top down, organizational leaders utilize open, two-way communication to develop a shared vision, giving workers a sense of meaning (Scholtes, 1998). Block (1993) proposed the term stewardship to describe a form of management that involved a redistribution of

purpose, power and privilege in the workplace. The idea that managers need to surrender the need to control and harboring an atmosphere which encourages and facilitates self-management is fundamental to this approach (Law, 2007). Kohn (1999) offered some suggestions to consider in place of performance appraisals if the goal of management is to foster improvement. He suggested a continuous process of two-way conversation between manager and employee which involves a change of ideas rather than judgments, and which is devoid of elements of ranking, competition, and compensation (Kohn, 1999). Kohn was particularly emphatic about severing any link between appraisals and compensation, arguing that such reward systems tend to decrease intrinsic motivation and diminish the notion of a task having meaning on its own merit (Law, 2007).

Management by objectives (MBO) is an alternate method of performance appraisal which is based on collaborative goal-setting from both employee and manager. This technique was first promoted in the 1950s by management theorist Peter Drucker (1954). MBO requires a manager and employee to agree upon specific, obtainable objectives with a set deadline (Drucker, 1954). For example, a sales manager may be required to increase his revenue by 25% within three months. Once this goal is set, the responsibility is on the sales manager to direct himself towards the objective. With the MBO technique, success or failure is easily defined (Drucker, 1954).

Lee, Chen, and Chang (2008) offered an approach called performance conversations as an alternative to appraisals. Under this approach, it is the responsibility of both manager and employee to maintain dialogue, seek solutions to challenges and trust each other (Lee et al., 2008). At the heart of this relationship should be an ongoing,

open, and honest solution-focused conversation which includes a system of communication, via a set of record keeping performance logs (Lee et al., 2008). These logs are intended to keep all parties on a track of open communication, and unlike management-recorded performance appraisals, both employee and manager are expected to record information to be shared with each other (Lee et al., 2008).

Despite the presence and availability of performance appraisal alternatives, many organizations continue to use traditional performance appraisals. Despite their omnipresence, there is a good deal of critique among practitioners in both the research and applied settings about how fair and effective performance appraisals have been historically and are in the current organizational climate.

Critique of performance appraisal. About seventy years ago, researchers began to investigate performance appraisals and the controversial issues surrounding them. Cronbach (1955) emphasized that researchers need to understand the processes by which they rate, and the biases and assumptions through which they filter information, to achieve effective ratings. Similarly, Landy and Farr (1980) pointed out that researchers have failed to explore issues involving raters themselves. Performance ratings are often challenged for their validity because “bias pervades the typical rating” (Wherry & Bartlett, 1982, p. 550).

The main criticism of performance appraisal is that they are inherently subjective, and contaminated by external artifacts (Landy & Farr, 1980). Instead of measuring ratees' performance, “ratings were stronger reflections of raters' overall biases” (Lance et al., 1994, p. 768). Another criticism is related to the assessment process, in which most

processes rely heavily on the rater conducting them. Even with the inclusion of a structured appraisal process and rating instrument, raters' personality characteristics and human judgment ultimately affect the performance ratings of a ratee (Tziner et al., 2005). From a cognitive perspective, rating errors are conceptualized to be the result of the limitations of human cognition (DeNisi, 1996), such as memory accessibility (Murphy & Balzer, 1986), cognitive style (Cardy & Kehoe, 1984; Härtel, 1993), and affect (Cardy & Dobbins, 1986). These approaches generally assume that raters involuntarily commit rating errors owing to either poor scale designs or to their own cognitive limitations. In response to performance appraisal, Likert (1959) summarized the experience as a negative one:

The aim of reviewing the subordinate's performance is to increase his effectiveness, not to punish him. But apart from those few employees who receive the highest possible ratings, performance review interviews, as a rule, are seriously deflating to the employee's sense of worth...not only is the conventional performance review failing to make a positive contribution, but in many executives' opinions it can do irreparable harm. (p.76)

Consistent with this view, other researchers have observed that employees and their supervisors often find appraisal both painful and de-motivating (Pfau, Kay, & Nowack, 2002). A survey of 2,004 employees was conducted in which the internal systems within organizations acknowledged to be intrinsic to organizational success were examined. Included in this examination, was the motivation system, where performance appraisal was a fundamental part of the process. Key findings showed that only 57% of

employees thought that their performance was rated fairly and that 60% of employees stated that they understood the measures used to evaluate their performance (Pfau et al., 2002).

The intimate nature of a feedback discussion paired with the possibility of delivering unsatisfactory news can be difficult for managers (Villanova et al., 1993). For example, Brett and Atwater (2001) found that recipients of negative feedback reacted to this feedback with anger and discouragement. Moreover, when compared to positive feedback, negative feedback was not regarded as useful or as accurate in nature (Spence & Keeping, 2009). By boosting ratings, raters can neutralize or avoid potentially uncomfortable situations. Due to the premise that negative consequences are usually the result of low ratings or negative ratings (i.e., being confronted about a low rating and having to communicate negative feedback), raters are usually expected to boost ratings in order to neutralize or avoid negative consequences (Spence & Keeping, 2009). Many organizations recognize this and offer training to equip managers with tools to engage in an effective and comfortable feedback session (Bernardin & Walter, 1977). For the same reasons, researchers have applied a great amount of effort to exploring feedback and how it affects ratings along with other segments of the performance management process (Villanova et al., 1993).

Another limitation of performance appraisal measurement is that it is difficult to obtain objective indices of performance for many job titles (Landy & Farr, 1980). In addition, personnel information may be applicable to a small portion of the employee

population in any organization. For example, 5% of the employees may have 100% of the accidents, or records are not well kept for all employees.

This distrust of the appraisal process is largely due to perceived links between appraisal and decisions about compensation and promotion. Despite the distrust and possible inefficiencies, the majority of organizations remain committed to the traditional supervisor evaluation approach.

Rater Bias in Performance Appraisals

All in all, there are many benefits to performance appraisals for employees, managers and organizations as a whole. These benefits include decisions about promotion and pay, employee development, assessment of potential and research development. However, the usefulness of performance appraisal is limited by subjectivity and rater bias. For an appraisal system to work, managers must understand their employees work well enough to appraise it, be trained for appraisal processes, and use appropriate and valid standards (Kramar et al., 1997). In reality, not every organization gives managers appraisal training to ensure that managers understand how to use the rating scale, the importance of the appraisals, and how to be aware of and address their own biases.

Raters do not function as neutral and objective observers of physical workplaces (Tziner et al., 2005). In reality, they are influenced by a variety of factors when it comes to giving ratings. Rater bias occurs whenever there is leniency, or harshness, present in the performance ratings. The most frequent problem that undermines the accuracy of performance appraisals is a form of rater bias in which raters assign ratings that are elevated relative to true performance levels (Ilgen & Feldman, 1983; Kane, Bernardin,

Villanova, & Peyrefitte, 1995). Among other detrimental effects, artificially-elevated ratings more quickly deplete the available funds designated for merit increases, produce more marginal rewards for ratees truly deserving better compensation, and generate perceptions of inequity (Kane et al., 1995). Because of these potential negative implications, rating elevation is likely the most significant form of rating distortion present in subjective ratings (Austin & Villanova, 1992).

Unintentional biases in ratings. A considerable amount of research on human information processing and cognition suggests that even when in a desirable climate and motivated to rate accurately, raters have highly imperfect perceptions and recall (Neisser, 1976). Researchers indicate that performance ratings are based on a cognitive categorization process (Bernardin & Beatty, 1984; Carroll & Schneier, 1982; Cooper, 1981; Feldman, 1981). According to their findings, at the time of a formal evaluation, raters recall the target person as belonging to an evaluative category to which the person previously had been assigned or processed. The person to be rated is then recalled, not as a composite of observed behaviors but as possessing characteristics that are generally representative of category members (Feldman, 1981). Thus, performance ratings are a function not only of observed behaviors but also of a category prototype, an abstraction based on the most common features of the category (Cantor & Mischel, 1977). Because performance appraisals are by nature inferential (i.e., a manager is never able to observe or recall all of his or her subordinates' behaviors), the accuracy of a performance rating will be a function not only of observed performance but also of the rater's sensitivity to the normative relationships among behaviors and his or her threshold for making

performance inferences (Nathan & Alexander, 1985). Raters who have little opportunity to observe behavior can still make accurate ratings if they are willing to make inferences about overall performance based on observed behavior (Nathan & Alexander, 1985). In contrast, raters with more extensive knowledge of employee behaviors can make inaccurate ratings if they make inaccurate inferences of observed behaviors (Nathan & Alexander, 1985).

Some of the most important conclusions to be drawn from research on human information processing are that our processing capabilities are limited and that perception and recall frequently do not match reality. The limitation on our processing capacity is handled by cognitive representations called schemata (Neisser, 1976). A schema directs our attention and aids in categorization and recall of information. However, a schema can also lead to systematic inaccuracies. Biased ratings may result when a rater relies on an irrelevant, over simplistic or otherwise faulty schema. Ratee traits and characteristics may elicit a schema which the rater employs to process and recall ratee performance (Bernardin & Cardy, 1981). For example, the gender of a ratee may be irrelevant to job performance but may set up a gender stereotype that may bias perception and recall of the ratee's performance. Research indicates that once a ratee is categorized, further perception and recall of that ratee's performance is biased toward that category or schema (Cantor & Mischel, 1977; Snyder & Swann, 1978). Additionally, humans are typically unaware of these biasing processes and will deny the operation of such a bias even when it is clearly present (Nisbett & Wilson, 1977).

These biases can be seen in the pattern of ratings, both within the rating forms for individuals and across ratings forms for different employees (Tsui & Barry, 1986). These within and across-form patterns are called halo and distributional errors, respectively. Halo is defined as a raters' failure to differentiate among different dimensions of the ratee's behaviors (Murphy & Balzer, 1989; Saal, Downey, & Lahey, 1980). In other words, halo effect is a rater's tendency to give similar ratings on all performance dimensions for a single ratee (Tsui & Barry, 1986), characterized as "inappropriate generalizations from one aspect of a person's performance on the job to all aspects of a person's job performance" (Latham & Wexley, 1981, p.255). Halo error may at least partially be due to the perceived similarity among rating categories (Cooper, 1981). For example, three categories may be rated similarly because the rater perceives them to be related even though the categories may, in reality, be independent. As measured by the magnitude of the intercorrelation among items obtained from each rating source, ratings by superiors consistently exhibit greater halo effects than self-ratings (Klimoski & London, 1974; Lawler, 1976; Parker, Taylor, Barrett, & Martens, 1958). This means that the presence of halo effect varies depending on whether the scenario is either top-down, in which a manager rates an employee or self-rating, where an employee rates him or herself. Another type of bias present in performance ratings is known as distributional error, which occurs when a rater tends to rate every participant the same. Distributional errors can include leniency, severity and central tendency errors. Leniency errors occur when the rater rates everyone at the desirable end of the scale while severity errors occur when the rater rates everyone at the undesired end of the scale (Bass, 1956; Hauenstein,

1992; McIntyre et al., 1984). Leniency can be described as a personal characteristic that leads an individual to consistently evaluate other people or objects in an extremely positive fashion. Central tendency errors occur when a rater rates everyone in the middle of the performance scale (Murphy & Balzer, 1989). As a result of distributional and halo biases, Murphy and Cleveland (1995) noted, “it is not unusual to find that 80% to 90% of all employees are rated as ‘above average’” (p. 275). Whisler (1958) suggested that the utility of performance appraisals as an accurate measure of employee performance is limited due to the hesitancy of raters to be completely honest for fear of negative repercussions. Harris (1994) argued that in order for raters to rate accurately they need to be motivated to do so. Predictors of rater motivations in Harris’ framework include such things as situational factors (e.g., being accountable to a supervisor), negative consequences (e.g., damage to subordinate-supervisor relationship), and rewards (e.g., likelihood of pay increases or promotions). A non-motivated rater exhibits less thorough and deliberate information processing techniques (Harris, 1994).

When raters possess cultural beliefs that are inconsistent with the practice of giving upward or lateral feedback, rating biases can be even more prevalent (Leslie, Gryskiewicz, & Dalton, 1998). This occurrence is especially important when using multi-source feedback, which is a practice that originated in the United States, and as such, is based on the assumptions of individualistic cultures and low-power distance values (Fletcher & Perry, 2001; Leslie et al., 1998; Shipper, Hoffman, & Rotondo, 2007). For example, providing objective feedback on an individual’s behaviors is based on individualistic values that emphasize personal striving and self-assertiveness (Morrison,

Chen, & Salgado, 2004). Moreover, the process of including peers and subordinates marks a redistribution of evaluating power and is by nature more compatible with low-power distance values that are less sensitive to status and hierarchy (Leslie et al., 1998; Shipper et al., 2007). This suggests that raters, especially peers and subordinates, may be more prone to rating biases when power distance and individualism-collectivism value orientations are inconsistent with multi-source feedback.

Intentional biases in ratings. Research has addressed the difficulty raters may have in maintaining their objectivity and neutrality (McGregor, 1957). The findings suggest that managers may possess a natural reluctance to rate their employees' performance because the performance appraisal process essentially asks managers to evaluate the worth of other human beings, a task with which the majority of people are probably uncomfortable with (McGregor, 1957). Most appraisal research has viewed rating inaccuracies as unintentional or unconscious mistakes (Cardy & Dobbins, 1986) and are based on the implicit assumption that raters are trying to rate accurately and that rating inaccuracies are a product of raters not having the skills, information, or tools necessary to rate accurately (Spence & Keeping, 2009). Contrary to these perspectives, researchers have suggested that rating errors may be the product of strategic decisions made by raters (Cleveland & Murphy, 1992; Kane, 1994; Murphy, Cleveland, Skattebo, & Kinney, 2004). Longenecker et al., (1987) interviewed executives and discovered that raters often knowingly give employees inaccurate performance appraisal ratings after deliberate consideration of the consequences of ratings. Evidence that rating inaccuracy has more to do with the deliberate, volitional distortion of performance ratings than was

previously recognized has been growing in recent years (Bernardin & Beatty, 1984; Bernardin & Villanova, 1986; Longenecker et al., 1987; Murphy & Cleveland, 1995). This notion is also supported by anecdotal evidence. For example, a survey of raters, ratees, and administrators of performance appraisal systems revealed that the majority of respondents in all these groups feel that rating inaccuracy stems much more from deliberate distortions than from raters' inadvertent, cognitive errors (Bernardin & Villanova, 1986). Additionally, researchers suggested that the act of rendering a performance rating is a motivated behavior and that managers do in fact rate in accordance with specific goals, such as a desire to motivate subordinates or concern with maintaining civil working relationships (Cleveland & Murphy, 1992). In other words, raters "have specific (and possibly) multiple goals in mind and they intend to provide ratings that are consistent with these goals" (Murphy et al., 2004, p. 158).

A considerable amount of theory and discussion exists as to what contextual factors raters consider when rating and what factors motivate a rater to knowingly rate inaccurately (Larson, 1984; Villanova & Bernardin, 1989). A comprehensive review of the literature by Spence and Keeping (2009) revealed three main reasons for performance rating distortion: (a) avoidance of negative consequences (Bass, 1956; Bernardin & Beatty, 1984; Curtis, Harvey, & Ravden, 2005), (b) compliance with organizational norms (Bernardin & Beatty, 1984; Decotiis & Petit, 1978; Dipboye, 1985; Harris, 1994; Larson, 1984; Longenecker et al., 1987; Mohrman & Lawler, 1983; Tziner et al., 2005), and (c) pursuit of self-interest (Bass, 1956; Harris, 1994; Ilgen, Mitchell, & Fredrickson, 1981).

Empirical data has indicated that these deliberate rating distortions also occur because of supervisors' feelings of discomfort with the appraisal system and its outcomes, and reflect their conscious efforts to produce ratings that will achieve personal goals (Murphy & Cleveland, 1995; Murphy et al., 2004). Similarly, other researchers have demonstrated that even when performance appraisals are conducted, supervisors frequently avoid potentially aversive situations by inflating the scores of their subordinates (Longenecker et al., 1987) particularly when they will be required to give face-to-face feedback (Landy & Farr, 1983). This alternative approach to performance evaluation conceptualizes that a part of rating inaccuracy is, in reality, not related to rating error; but it is intentionally introduced by the rater to achieve specific goals in organizational contexts (Cleveland & Murphy, 1992; Murphy & Cleveland, 1991, 1995; Murphy et al., 2004; Wong & Kwong, 2007). For example, raters pursuing a harmony goal will increase their mean ratings and will decrease their rating differentiation, and raters pursuing a fairness goal will inflate their mean ratings and decrease the rating differentiation (Wong & Kwong, 2007). These studies suggest that performance evaluation is not just a measurement process, but it is also a social process and a communication process. In other words, raters are not passive participants in the process but are active participants with the ability and motivation to distort ratings intentionally to attain predetermined goals (Wang, Wong & Kwong, 2010).

The notion that supervisors are not entirely objective when rating their employees' performance is not surprising when considered in the context of the consequences of rating decisions. Formal performance appraisal systems are used in

about 90 percent of organizations (Bernthal et al., 1997) for administrative decisions such as promotions and terminations, as well as for employee development (Murphy & Cleveland, 1995). As a result, performance appraisals often directly affect employee development, career trajectories and the allocation of money and resources. Due to the major significance of appraisal ratings, both personally and professionally, it is understandable how raters might have a difficult time maintaining objectivity when rating their employees (Spence & Keeping, 2009). Like the main effects of rater goals on ratings, research on personality characteristics have also showed a tendency to distort ratings (Tziner et al., 2005). Some researchers believe that raters have the ability to rate fairly, but can choose not to, given the context of their situation (Cleveland & Murphy, 1992). Other research posits that raters can be heavily influenced by non-performance factors such as personality variables and cannot control for this bias without training (Bernardin, 1978). Not surprisingly, several researchers have found that if given the option, many supervisors would choose to not give performance feedback to their subordinates, especially if the subordinate has performed poorly (Fried, Tiegs & Bellamy, 1992).

Predictors of Performance

A number of models of performance have suggested that the apparent shortcomings of performance ratings are the result of rater individual attributes. However, the research on individual factors provides relatively few general conclusions. Most studies examine only one or a few characteristics, so it is likely that unmeasured variables may have some effect on the results of any single study. Nevertheless, previous

research provides some information on how performance ratings are influenced by the effects of individual factors.

One of the earliest and most widely-researched demographic variables in performance appraisal bias has been rater gender (Landy & Farr, 1980). Since the 1970s, research has been inconclusive and inconsistent with respect to rater gender and its impact on performance appraisal, despite obtaining performance appraisal data in various contexts, including instructional settings (Elmore & LaPointe, 1974), simulated work settings (Rosen & Jerdee, 1973), and laboratory research settings (Jacobsen & Effertz, 1974). In a simulated work setting, London and Poplawski (1976) found that female raters gave higher ratings on some dimensions but not on overall performance on simulated appraisal and interview situations. Similarly, another group of researchers found that females gave higher ratings than did males when evaluating performance in a simulated work setting, especially for high levels of performance (Hamner, Kim, Baird, & Bigoness, 1974). Another dimension that has been researched is the ratee's gender, also known as demographics-based rater bias (Arvey & Murphy, 1998). This phenomenon occurs when employees with certain demographic characteristics receive systematically lower or higher appraisal ratings (Arvey & Murphy, 1998). For example, studies have shown that managers' performance appraisals are sometimes influenced by the gender, race, or age of the ratee (Murphy & Cleveland, 1995). The issue of rater and ratee gender is still of particular importance because biased ratings like this result in discrimination (Demuijnck, 2009) and inequity (Ngo, Foley, Wong, & Loi, 2003).

Researchers have also explored if rater performance appraisal experience has an effect on ratings, but results of these studies are mixed (Landy & Farr, 1980). Jurgensen (1950) found that more experienced raters had more reliable ratings. Related to this idea of experience is the topic of expertise. A large body of literature has examined the performance differences between experts and novices on complex tasks (Chase & Ericsson, 1982; Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981). Across a variety of task domains, research has shown that individuals with more expertise have a larger knowledge base in their area of expertise compared to novices. More importantly, this knowledge base is organized into meaningful schemas, such that larger units of information are constructed from meaningful relations among smaller units (Chase & Simon, 1973). This means that experts are better able to map new stimuli from a relevant domain onto existing knowledge structures in that same domain. Extrapolating the literature on expertise to appraisal experience, it seems reasonable to expect that experienced raters will have richer knowledge structures regarding performance appraisals. So, when presented with non-performance influence to inflate ratings, raters should be able to incorporate these variables into their ratings without being distracted by them. To support this, Spence and Keeping (2009) found that individuals with more years of performance appraisal experience provided lower performance ratings than those with less appraisal experience. Conversely, Mandell (1956) noted that raters with more than four years of experience as supervisors tended to be more lenient in their ratings than were raters with less experience. Cascio and Valenzi (1977) found a significant effect of rater experience, but noted that it accounted for only a small percentage of total rating

variance. While in another study, rater experience had no significant effect on ratings (Klores, 1966). Due to inconsistent results, this demographic variable requires further examination to determine how far-reaching its effect on appraisals may be. According to Landy and Farr (1980), rater experience appears to positively affect the quality of performance ratings, but the mechanism responsible (e.g., more appraisal training, better observation skills, better knowledge of job requirements, etc.) is not known. Rater experience seems to be a potentially important demographic variable to examine when discussing the effects of non-performance variables on performance appraisals.

While many individual factors have yielded mixed results when used as predictors for performance ratings, reviews of the research on the connection between broad personality characteristics and behavior in organizations (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991) confirm the relevance of several aspects of personality for understanding job performance, interaction in groups and other organizational phenomena. Other research demonstrates that raters' attitudes, beliefs, personalities and orientations toward performance systems intrude on rating behavior (Tziner et al., 2005). From this perspective, aspects of rating bias can be driven by stable rater tendencies and personality traits may be effective in explaining these tendencies (Kane et al., 1995; Borman & Hallam, 1991). Kane et al., (1995) found a mean stability coefficient of .48 across three studies. Villanova et al., (1993) reported a mean stability coefficient of .63 for students evaluating their peers on group projects. These findings suggest that rating elevation might be predicted using measures of individual differences. The current study

aims to investigate four factors that have been shown to be associated with performance ratings.

Conscientiousness. Conscientiousness is one of the five dimensions of the Five Factor Model (FFM) of personality, (McCrae & John, 1992). The Five Factor Model of personality describes an individual's personality as being largely explained by five independent factors and has gained widespread acceptance by personality researchers and has greatly influenced the research on individual differences (Barrick & Mount, 1991; Goldberg, 1993; Salgado, 1998). The five factors of the FFM are openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. Individuals scoring high on conscientiousness strive for excellence, are characterized as hard-working, have high performance standards, set hard-to-accomplish goals (Costa & McCrae, 1992). The current study will be using this definition of conscientiousness according to Costa and McCrae (1992). Like all dimensions on the FFM, according to both the Costa and McCrae NEO PI-R and Goldberg NEO-IPIP models, conscientiousness is considered to be a continuous dimension of personality.

The development of the five factor model is based on trait theory, which states that traits are relatively stable over time, differ across individuals and directly influence behavior (Kassin, 2003). There are an infinite number of potential traits that can be applied to personality, but through the statistical technique of factor analysis, clusters of traits have shown to reliably correlate together. Costa and McCrae (1992) developed the FFM by identifying five independent personality traits that remain stable over time and influence behavior in a wide variety of situations.

Conscientiousness has been linked to several positive outcomes across educational, health, and personnel psychology, and appears to be the personality trait with the most predictive utility (Barrick & Mount, 1991; Bogg & Roberts, 2004; Poropat, 2009). The attribute of conscientiousness has proved to be especially useful for predicting a range of behaviors in organizations (Barrick & Mount, 1991), and it is likely to be a particularly important factor in shaping performance appraisal behavior (Tziner et al., 2005; Bernardin et al., 2000). Past research has predicted, and found, that raters high in conscientiousness are less prone to rating elevation (Bernardin et al., 2000). As suggested by Bernardin et al. (2000), raters who were less conscientious were more lenient and less accurate in their ratings than those who were more conscientious. Many researchers have proposed that the predictive power of dispositional factors such as rater conscientiousness may be moderated by accountability factors related to the characteristics of the performance appraisal system (Tziner et al., 2001). For example, Bernardin et al. (2000) argued that their findings may have been unique to a rating scenario in which ratings had administrative significance. Raters expected that their ratings were attached to them personally or that raters anticipated future interaction with the ratees receiving feedback on their performance (Bernardin et al., 2000). Using Wright and Mischel's (1987) conditional view of dispositional constructs as applied to rating behavior, Bernardin et al. (2000) proposed that if any accountability factor were to be absent in an appraisal situation, then the predictability of conscientiousness on to a rating level may be lessened. Harris (2006) found that raters who were relatively low on conscientiousness tended to rate leniently, but only when their ratings were directly linked to them while

raters who were relatively high on conscientiousness did not inflate ratings when they were identified.

These findings suggest that being identified as a rater might cause potential discomfort which then leads conscientious raters to rate more leniently (Harris, 2006). The conscientiousness trait has been shown to determine the degree to which an individual is willing to direct efforts to a task even when there is considerable external pressure and motivation to discontinue efforts (Costa & McCrae, 1992). One such notable motivation to curb performance appraisal efforts is rater discomfort. It is well established that the discomfort associated with assigning a performance rating motivates individuals to inflate ratings in order to avoid feelings of discomfort (Villanova et al., 1993). In such case, the rater compromises the validity of the performance rating by engaging in a thought process that devalues the importance of the appraisal task. As such, raters high in conscientiousness are less willing to give in to the pressures associated with rater discomfort.

Agreeableness. Agreeableness is another trait from the FFM that has been examined in relation to performance ratings (Bernardin et al., 2000; Tziner et al., 2001; Yun, Donahue, Dudley, & McFarland, 2005). Most agreeable individuals are more trustful, sympathetic, cooperative, and polite (Costa & McCrae, 1992). However, individuals showing high scores on agreeableness may also be more sympathetic, cooperative, dependent and self-effacing, with an orientation toward agreement and acquiescence (Costa & McCrae, 1992). Subsequent research shows that highly agreeable individuals also tend to manifest a strong desire for social approval, value relationships

(Meier, Robinson, & Wilkowski, 2006) and avoid social conflict (Jensen-Campbell, Knack, Waldrip, & Campbell, 2007) where as their highly conscientious counterparts focus more on tasks. Thus, highly agreeable individuals should be more lenient in their performance appraisals, particularly when there is significant ongoing or anticipated social interaction between raters and ratees and there is potential for confrontation over the ratings (Bernardin et al., 2000). The current study will be using this definition of agreeableness according to Costa and McCrae (1992). According to both the Costa and McCrae NEO PI-R and Goldberg NEO-IPIP models, agreeableness is considered to be a continuous dimension of personality, rather than a categorical type of person. As with the conscientiousness domain, the identification of the agreeableness domain is based on trait theory, which states that traits are relatively stable over time, differ across individuals and directly influence behavior (Kassin, 2003).

Past research has predicted, and found, that raters high in agreeableness tend to produce more elevated ratings than those low in agreeableness (Bernardin et al., 2000). Individuals high in agreeableness may produce more lenient ratings in particular situations, such as when raters anticipate future interaction with the ratees or when the raters are solely responsible for the ratings (Kane et al., 1995). Similarly, Yun et al., (2005) found that when participants expected to have a face-to-face feedback meeting, highly agreeable raters produced higher performance ratings than those lower on agreeableness.

These findings suggest that an individual who is high in agreeableness is more likely to give in to social pressures that would motivate the assignment of inflated ratings.

The act of performance appraisal can be a source of discomfort for those who are highly agreeable as compared to others because of the potential social repercussions associated with performance ratings and feedback (Villanova et al., 1993). An individual who is high in agreeableness is very much concerned with being generous, kind, and sympathetic and is likely to be lenient in evaluations to satisfy that concern feedback (Villanova et al., 1993). Therefore, highly agreeable individuals are likely to experience increased feelings of rater discomfort.

Rater discomfort. Rater discomfort occurs when an employee experiences uneasiness, heightened anxiety or withdrawal when charged with rating another individual for the purposes of a performance appraisal (Villanova & Bernardin, 1989). The current study will use this definition when examining rater discomfort.

The concept of rater discomfort was born from job compatibility theory. Job compatibility theory refers to the extent to which employees maintain preferences for job characteristics that are consistent with the actual demands of the job (Villanova et al., 1993). According to the job compatibility framework, employees whose preferences are at odds with their job characteristics tend to report greater discomfort in performing job activities and manifest behaviors indicative of less job involvement and higher withdrawal and avoidance (Villanova et al., 1993; Spence & Keeping, 2009). Conducting performance appraisals is not a job demand that is typically consistent with employee preferences and job characteristics.

Researchers have found that raters frequently report discomfort with several facets of the performance appraisal process, including the monitoring of subordinates'

performance, evaluating employee performance, and providing performance feedback (Murphy & Cleveland 1991). Villanova et al., (1993) noted that raters who show high levels of appraisal discomfort are more likely to provide inflated ratings and are less likely to distinguish among employees. In other words, raters may be motivated to assign uniformly high ratings in order to avoid discomfort associated with making difficult judgments of others' performance. Rater discomfort with performance appraisals has been found to be positively associated with rating leniency. Villanova et al., (1993) developed a performance appraisal discomfort scale and found a positive correlation between raters' discomfort levels and performance ratings.

These results are consistent with the job compatibility framework, which purports that discomfort is produced when an employee's preferences are in conflict with the requirements of his or her job. These feelings of discomfort are thought to decrease involvement and lead to withdrawal behavior in the employee (Spence & Keeping, 2009). In the case of performance appraisals, raters are thought to withdraw from the task of providing representative performance ratings by providing their employees with uniformly high ratings. This research implies that raters may alter performance ratings as a type of preventive behavior.

Self-efficacy. Self-efficacy is an individual's belief in his or her own ability to perform an action or task successfully, to meet the demands of a given situation (Wood & Bandura, 1989). Individuals differ in self-efficacy, or the extent to which they believe they have the information, tools, and skills necessary to perform a task competently. Self-efficacy, as perceived by the individual, is likely to play a motivational role and to

influence behavioral choices, affecting the mobilization of efforts and the perseverance with which goals are pursued. Self-efficacy can be described as a holistic self-assessment based on an individual's cumulative experiences. The current study will be using this definition of general self-efficacy according to Wood & Bandura (1989).

Previous research has found that situational self-efficacy predicts several important work-related outcomes, including job attitudes (Saks, 1995), training proficiency (Martocchio & Judge, 1997), and job performance (Stajkovic & Luthans, 1998). In addition to the relationships observed in previous research between situational self-efficacy and a variety of performance indicators, researchers observed a more stable and trait-like, general dimension of self-efficacy, which has since been termed general self-efficacy (GSE) (Gardner & Pierce, 1998; Judge, Erez, & Bono, 1998; Judge, Locke, & Durham, 1997).

The concept of self-efficacy is the basis of Bandura's social cognitive theory which explains that an individual's behaviors and thought processes are influenced by the previous actions that the individual has performed or observed (Bandura, 1977). Furthermore, social cognitive theory posits that social experiences and observational learning influence the development of the self-efficacy trait that is then applied to all future tasks and experiences (Bandura, 1988). Therefore, self-efficacy is not just an overall self-appraisal, but is a trait that determines future behavior, with those that are higher in self-efficacy less likely to avoid a task that is perceived to be difficult or aversive.

Researchers have posited that self-efficacy may be important when it comes to producing less inflated ratings and more positive attitudes toward the appraisal process (Bernardin & Beatty, 1984; Bernardin & Buckley, 1981; Napier & Latham, 1986; Tziner, 1999; Villanova et al., 1993). In the context of performance management, specific self-efficacy pertains to a rater's belief that he or she may orchestrate performance in the course of fulfilling their role obligation as it pertains to performance management (Bernardin & Villanova, 2005). According to this definition, raters with low self-efficacy might lack sufficient motivation to provide well-documented, solidly grounded, reliable, and accurate evaluations (Frayne & Latham, 1987). Research shows that specific self-efficacious raters believe themselves capable of executing socially-demanding behaviors that have important consequences for their relationship with ratees (Bernardin & Villanova, 2005). Self-efficacy training, directed at increasing rater confidence and capability in identifying particular performance levels and providing negative feedback to performer, has been associated with decreased rater discomfort in the appraisal process and as a result decreased levels of rating bias caused by the discomfort (Bernardin & Villanova, 2005). The thinking is that self-efficacy training in the appraisal context should facilitate more confidence and capability in raters for providing negative feedback to performers. Traditional rater training focuses on increasing rater self-efficacy for the behavioral competencies of collecting more relevant observations, avoiding rater biases. However, the acquisition of these skills, albeit relevant for performing the task of performance appraisal, may not be sufficient to offset rater self-doubt in conducting other relevant behaviors, such as resolving rating disputes that might arise in appraisal

interviews (Neck, Stewart, & Manz, 1995). Rater self-efficacy in handling such interpersonal demands appears necessary to offset rater motives to avoid disputes and the attenuating effect they have on rater training aimed at improving accuracy. Accordingly, a more comprehensive rater-training program provides raters with skills that span the rating process, from observation to feedback (Hauenstein, 1992).

The effectiveness of self-efficacy training in minimizing rater discomfort displays the importance of self-efficacy when it comes to bias in appraisal, especially since self-efficacy is a more dynamic trait which can be somewhat changed and improved upon through awareness and training. This link between rater discomfort and self-efficacy is an important finding because it gives us a behavioral explanation for bias in appraisal.

The level of performance. If raters are not capable of compensating ratings to account for the influence of situational factors on observed performance, then the resulting ratings will be contaminated with situational influences and will be unlikely to reflect the true level of performance in a valid manner. Because of the important implications performance evaluations hold (personnel decisions, detection of employee performance) it is important to examine if raters consider situational influences when evaluating employee performance. Research examining the relationship between actual performance and performance ratings has been dominated by experimental studies that have generally found a significant relationship between actual performance and performance ratings (Bigoness, 1976; DeNisi & Stevens, 1981; Grey & Kipnis, 1976; Hamner et al., 1974; Jones, Shaver, Goethals, & Ward, 1968). In one study, 20 MBA student participants were asked to evaluate production environment scenarios (Carson,

Cardy, & Dobbins, 1991). Participants were provided with descriptions of ability, effort, time to setup production run, difficulty and observed performance in the form of production data. They then evaluated fictitious employees on a 7-point scale. Results showed that 70.9% of the variance in performance ratings was a function of actual productivity. Productivity data dominated descriptions of system factors in influencing performance ratings. As a result, Carson et al., (1991) concluded that their results appear to support Deming's (1986) criticism that raters are incapable of considering the influence of situational factors and will inappropriately attribute variation in performance to the individual employee. This study supports the notion that raters do not consider information when evaluating performance. On a similar note, the results of a past study suggest that actual performance accounts for the most variance on raters' subjective performance evaluations (Huber, Neale, & Northcraft 1987). Using city government managers as raters, Huber et al., (1987) found that objective performance accounted for the largest amount of variability in judgmental performance ratings; however, rater characteristics (e.g., sex, age, and rater experience) moderated the relationship between objective performance and rater judgments. According to Huber et al., (1987) raters relied on heuristics as they processed performance appraisal information (Landy & Farr, 1983) to simplify the cognitively complex task; therefore, some ratee information inappropriately influenced the raters' judgments. Alternatively, Scullen, Mount, and Goff (2000) found that actual ratee performance accounted for only 30% of total variance. In this study, two large data sets, consisting of managers who received developmental ratings on performance dimensions were used. The results showed that idiosyncratic rater

effects accounted for over half of the rating variance in both data sets, while the effect of ratee performance was less than half the size of the idiosyncratic rater effects. This finding means that what is being rated does not account for more variance than who is doing the rating. When comparing the varying results of the two aforementioned studies, it is difficult to assess the role of the level of performance and know how strongly it may influence performance ratings.

Summary and Conclusion

Because of its many implications for employees and their organizations, accurate assessment of employee job performance continues to be a topic of great interest to organizational researchers. Subjective ratings of job performance inform promotion, training and development, transfer and termination decisions, are common to 90% of organizations (Bernthal et al., 1997).

Nevertheless, the material and psychological significance of ratings for employees elevate the appraisal process to one of considerable apprehension and drama for raters (Kozlowski, Chao, & Morrison, 1998). This negative aspect of performance appraisal has even led some to advocate abolishing formal appraisals altogether (Coens & Jenkins, 2002). Alternative sources of appraisal data do exist that may produce more objective results, like quantity and quality of production or service, but these indices are unavailable as measures of performance for the majority of jobs (Bernardin & Villanova, 2005).

A wealth of research exists studying the psychology of appraisal and feedback processes, focusing on individual rater differences, rater discomfort, rater motivation, and

employee performance (Fletcher, 2001; Smither, London, & Richmond, 2005). However, few studies have gone beyond focusing on a few individual factors, and a sufficient investigation into the interplay among predictive factors has not yet been achieved.

CHAPTER III

The Present Study

This chapter describes information pertinent to the objectives of the present study. It begins with a presentation of the importance and the purpose of the study. The chapter concludes with the research hypotheses and supporting rationale.

Importance of the Study

In many organizations, performance appraisals remain a paradox of effective human resource management. Under ideal circumstances, performance appraisals assess individual performance objectively and fairly, and can provide valuable performance information to a number of critical human resource activities, such as the determination of training needs, selection criteria, performance documentation (Cleveland et al., 1989), and employee performance feedback (Ilgen et al., 1979). In reality, managers are inconsistent in applying objective criteria to performance appraisals, resulting in unreliable and sometimes deliberately-distorted evaluations (Folger et al., 1992). The undesirable effects of poorly-performed performance evaluations include employee disengagement, promotions and rewards being inappropriately allocated, an inability to accurately identify high-performing employees, and lost revenue related to these detrimental effects.

Furthermore, the material and psychological significance of ratings for employees elevate the appraisal process to one of considerable apprehension and drama for raters (Kozlowski et al., 1998). Although subjective appraisals are common to 90% of organizations (Bernthal et al., 1997), they are still steeped in controversy. There is

evidence that appraisal systems are a practical challenge to those who design them and to the managers and employees who must use them. As Banks and Murphy (1985) noted:

Organizations continue to express disappointment in performance appraisal systems despite advances in appraisal technology. Appraisal reliability and validity still remain major problems in most appraisal systems, and new appraisal systems are often met with substantial resistance. In essence, effective performance appraisal in organizations continues to be a compelling but unrealized goal. (p. 336)

The controversy over appraisal has even led some to advocate abolishing formal appraisals altogether (Coens & Jenkins, 2002). At the same time, alternatives to subjective appraisal are not very appealing. Alternative sources of appraisal data do exist that may produce more objective results, like quantity and quality of production or service, but these indices are unavailable as measures of performance for the majority of jobs (Bernardin & Villanova, 2005). In addition, it is common knowledge that these more objective indices are notoriously deficient and prone to contamination (Austin & Villanova, 1992).

Therefore, the identification of the drivers of these inefficiencies is paramount to organizations' ability to yield the maximum potential of employee performance appraisal. The practical implications of a better understanding of undesired variability in performance appraisal are particularly important because performance appraisal is a process which can be expensive, resource-intensive, and is popularly-perceived as ineffective. Although performance appraisal research is quite vast and extensively

explains the effects of individual predictors of performance ratings, it fails to examine the interaction among these predictors, and ignores the effects of rater characteristics on performance ratings.

The present study contributes to academic literature and applied settings in three ways. First, it aims to identify a number of predictors of performance ratings, such as conscientiousness, agreeableness, self-efficacy in performance appraisals, and rater discomfort in an attempt to understand factors that trigger distortion in performance appraisal. It also aims to examine the mediating role of rater discomfort on the relationship between conscientiousness and the level of performance ratings. As well as on the relationship between agreeableness and the level of performance ratings. In addition, the current study intends to explore the mediating role of rater discomfort on the relationship between self-efficacy in performance ratings and the level of performance ratings. It also aims to examine the moderating influence of performance level on the relationship between rater discomfort and the level of performance ratings. The present study is important because the mediating and moderating relationships among these key variables have not been examined in a single study before. Furthermore, the results of this study can assist organizations to administer specific appraisal training in an attempt to overcome rating bias, triggered by the key variables in this study. The results of the study can also assist in justifying the development of employee training, such as rater self-efficacy training for raters, to eliminate this bias. Last, it can lead to increased trust and faith in company appraisal processes which in turn may increase employee motivation.

Purpose of the Study and Hypotheses

The purpose of the present study is to examine the causal relationship between the level of performance and the level of performance ratings. Additionally, the moderating effect of the level of performance on the relationship between rater discomfort and the level of performance ratings will be examined. Also, the current study seeks to examine the mediating role of rater discomfort on the relationship between conscientiousness and the level of performance ratings, as well as on the relationship between agreeableness and the level of performance ratings, and on the relationship between self-efficacy in performance ratings and the level of performance ratings. Lastly, the present study seeks to examine the moderating role of conscientiousness on the relationship between rater discomfort and performance rating.

To test these mediating and moderating relationships, I designed a field experiment in which the level of performance of an employee is manipulated. More specifically, participants will be asked to read one of two different fictitious scenarios that describe a high-performing or a low-performing employee (manipulation of the between-subjects independent variable called the level of performance), and then asked to provide performance ratings for this employee (called the level of performance rating as the dependent measure). The participants will then be asked to complete a questionnaire that assesses their level of conscientiousness, agreeableness, self-efficacy, and rater discomfort.

Researchers have found that raters frequently report discomfort with several facets of the performance appraisal process, including the monitoring of subordinates'

performance, evaluating employee performance, and providing performance feedback (Murphy & Cleveland, 1991). Villanova et al., (1993) noted that raters who show high levels of appraisal discomfort are more likely to provide inflated ratings and are less likely to distinguish among employees. In other words, raters may be motivated to assign uniformly high ratings in order to avoid discomfort associated with making difficult judgments of others' performance. Rater discomfort with performance appraisals has been found to be positively associated with rating leniency (Villanova et al., 1993). Villanova et al., (1993) developed a performance appraisal discomfort scale and found a positive correlation between raters' discomfort levels and performance ratings.

Research examining the relationship between the level of performance and subjective ratings of that performance has been dominated by experimental studies that have generally found a significant relationship between actual performance and performance ratings (Bigoness, 1976; DeNisi & Stevens, 1981; Grey & Kipnis, 1976; Hamner et al., 1974; Jones et al., 1968). Accordingly, the following hypothesis was formed.

Hypothesis 1. When the level of performance indicates that the actual performance is high, performance ratings will be higher (See Figure 1).

When the level of performance reports that performance is low, there will be a positive relationship between rater discomfort and the level of performance ratings. In other words, raters who feel high discomfort in making evaluations, when tasked with evaluating a low performing employee, will likely make inflated evaluations. When the level of performance indicates that actual employee performance was high, however,

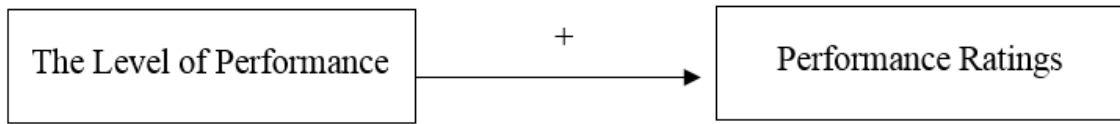


Figure 1. The predictive relationship between the level of performance and performance ratings.

there is no relationship between a raters' discomfort and the level of performance ratings. In other words, rater discomfort will not have any effect on the level of performance ratings assigned to the high-performance case. Accordingly, the following hypothesis was formed.

Hypothesis 2. The level of performance moderates the relationship between rater discomfort and the level of performance rating such that when the level of performance about employee performance indicates that his/her performance is actually low, there will be a positive relationship between rater discomfort and performance ratings. On the other hand, when the level of performance about employee performance indicates that his/her performance is actually high, this relationship will disappear (See Figures 2 and 3).

As reviewed in the literature review section above, conscientiousness determines the degree to which an individual is willing to direct efforts to a task (Costa & McCrae, 1992). Furthermore, because conscientious individuals tend to be more task focused, they are motivated by performing the task at hand, even when there is considerable external pressure and motivation to discontinue (Costa & McCrae, 1992). One such notable motivation to curb performance appraisal efforts is rater discomfort. It is well established that the discomfort associated with assigning a performance rating motivates individuals to inflate ratings in order to avoid feelings of discomfort (Villanova et al., 1993). In such case, the rater compromises the validity of the performance rating by engaging in a thought process that devalues the importance of the appraisal task. Thus, raters high in conscientiousness are less willing to give in to the pressures associated with rater discomfort because the context of performance appraisal possesses specific features that

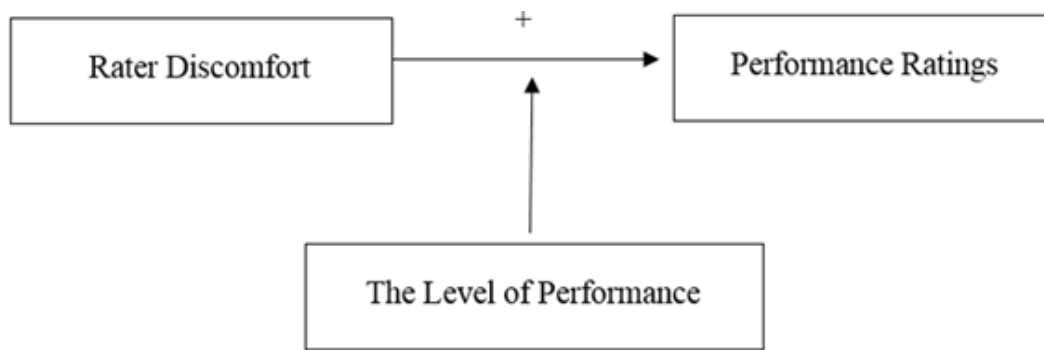


Figure 2. The research model for Hypothesis 2.



Figure 3. Moderational effect of the level of performance about performance on the relationship between rater discomfort and the level of performance rating.

are relevant to the trait of conscientiousness, such as accountability and the administrative significance of the ratings. Accordingly, the following hypothesis was formed.

Hypothesis 3. Rater discomfort mediates the relationship between conscientiousness and the level of performance ratings such that high levels of conscientiousness predict low levels of rater discomfort, which in turn predict less inflated performance ratings (See Figure 4). Individuals high in conscientiousness experience less rater discomfort because they actively seek information about the performance level, this in turn makes them better prepared to make performance ratings. Specifically, when information indicates low performance, discomfort ratings will be higher. Also, when information indicates high performance, there will be no relationship.

The act of performance appraisal can be a source of discomfort for those who are highly agreeable as compared to others because of the potential social repercussions associated with performance ratings and feedback (Villanova et al., 1993). An individual who is high in agreeableness is very much concerned with being generous, kind, and sympathetic and is likely to be lenient in evaluations to satisfy that concern feedback (Villanova et al., 1993). Therefore, highly agreeable individuals are likely to focus more on the relationship aspect rather than the task at hand and experience increased feelings of rater discomfort. Moreover, agreeable people enjoy being liked by others and seek opportunities when that can happen. Seeking out these opportunities conflicts with formal judging another individual's performance and makes highly agreeable individuals even more uncomfortable. This idea of wanting to be liked by others can be supported by a

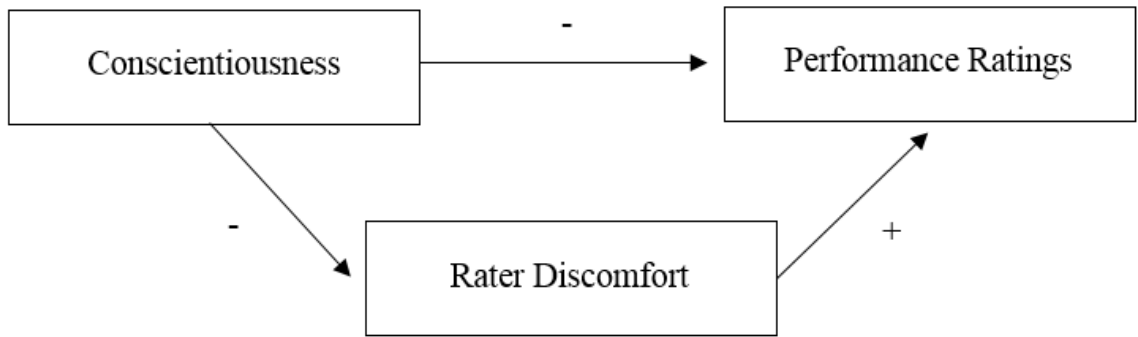


Figure 4. Research model for Hypothesis 3.

need for acceptance and fear of social rejection. The need for love and belongingness is a fundamental human motivation (Maslow, 1954). People have a strong drive to form and maintain caring interpersonal relationships and need both stable relationships and satisfying interactions with the people in those relationships (Baumeister & Leary, 1995). If either of these two ingredients is missing, people will begin to feel lonely and unhappy, and thus rejection is a significant threat. This fear of rejection can lead to conformity or normative influence and compliance to the demands of others (Williams & Zadro, 2001). Accordingly, the following hypothesis was formed.

Hypothesis 4. Rater discomfort mediates the relationship between agreeableness and the level of performance ratings such that high levels of agreeableness predict high levels of rater discomfort, which in turn predict higher performance ratings (See Figure 5). When information indicates low performance, discomfort ratings will be higher. However, when information indicates high performance, discomfort ratings will be lower.

Self-efficacy can be defined as “beliefs in one’s capability to mobilize the motivation, cognitive resources, and courses of action needed to meet given situational demand” (Wood & Bandura, 1989, p. 408). Relatedly, there has been a focus on general self-efficacy (GSE) in the past 20 years, which is a more trait-like generality dimension of self-efficacy (Judge et al., 1997). GSE can be defined as “individual’s perception of their ability to perform across a variety of different situations” (Judge et al., 1998, p. 170). GSE captures differences among individuals

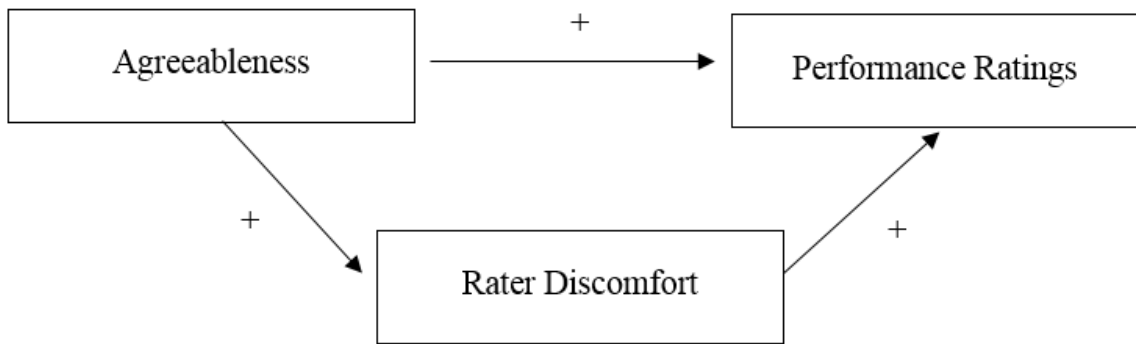


Figure 5. Research model for Hypothesis 4.

in their tendency to view themselves capable of meeting task demands in a wide variety of contexts. Specific self-efficacy (SSE) is a construct that grew out of GSE, and focuses on a task-specific state (Gist & Mitchell, 1992). Several researchers have suggested that GSE is a motivational trait and SSE is a motivational state (Gardner & Pierce, 1998; Judge et al., 1997). Although, both include beliefs about one's ability to achieve desired outcomes, the constructs differ in scope (general versus task-specific). GSE has been found to moderate the impact of external influences like performance feedback and training, for example (Eden, 2001). However, despite a large amount of empirical research, there are many criticism of GSE, mainly that the utility of GSE for both practice and theory is low (Stajkovic & Luthans, 1998). Bandura (1997) argued that GSE measures have "no relation to efficacy beliefs related to particular activity domains" (p. 42). As a result, a new general self-efficacy scale (NGSE) was developed and found to predict SSE (Chen et al., 2001). The current study will be measuring new general self-efficacy (NGSE). Self-efficacy training, directed at increasing rater confidence and capability in identifying particular performance levels and providing negative feedback to performer, has been associated with decreased rater discomfort in the appraisal process and as a result decreased levels of rating bias caused by the discomfort (Bernardin & Villanova, 2005). Therefore, low self-efficacious raters are more likely to feel higher levels of rater discomfort. Accordingly, the following hypothesis was formed.

Hypothesis 5. Rater discomfort mediates the relationship between self-efficacy for performance appraisals and the level of performance ratings (See Figure 6). When

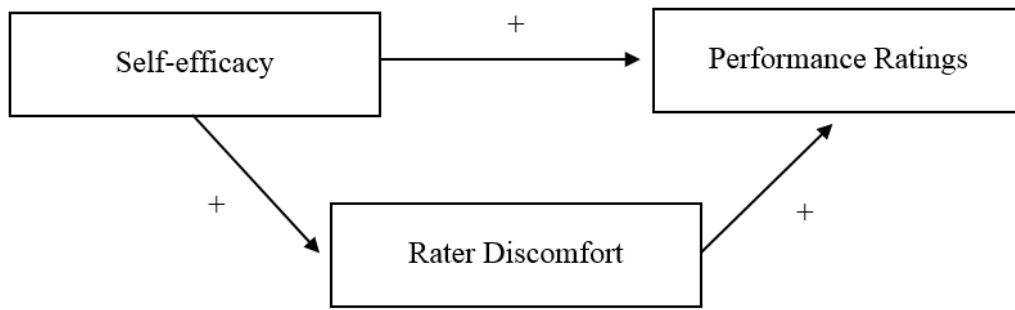


Figure 6. Research model for Hypothesis 5.

information indicates low performance, discomfort ratings will be higher. However, when information indicates high performance, discomfort ratings will be lower.

Hypothesis 6. Conscientiousness moderates the relationship between rater discomfort and performance ratings (See Figures 7 and 8). Rater discomfort will be more strongly associated with increased performance ratings under conditions of low conscientiousness and information indicating a low performance level, as opposed to a high performance level.

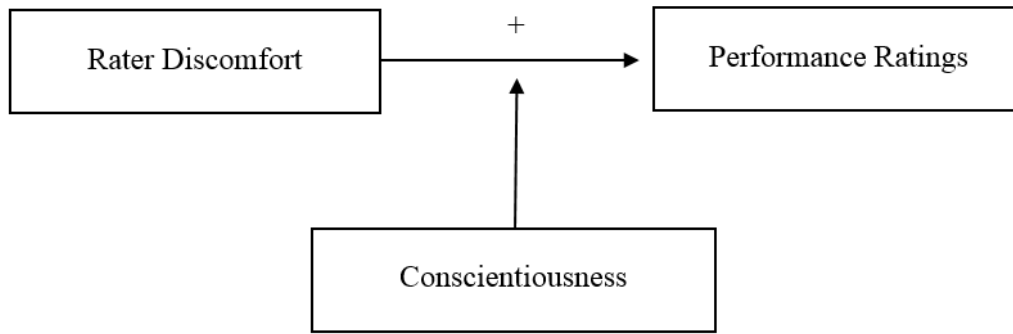


Figure 7. Research model for Hypothesis 6.

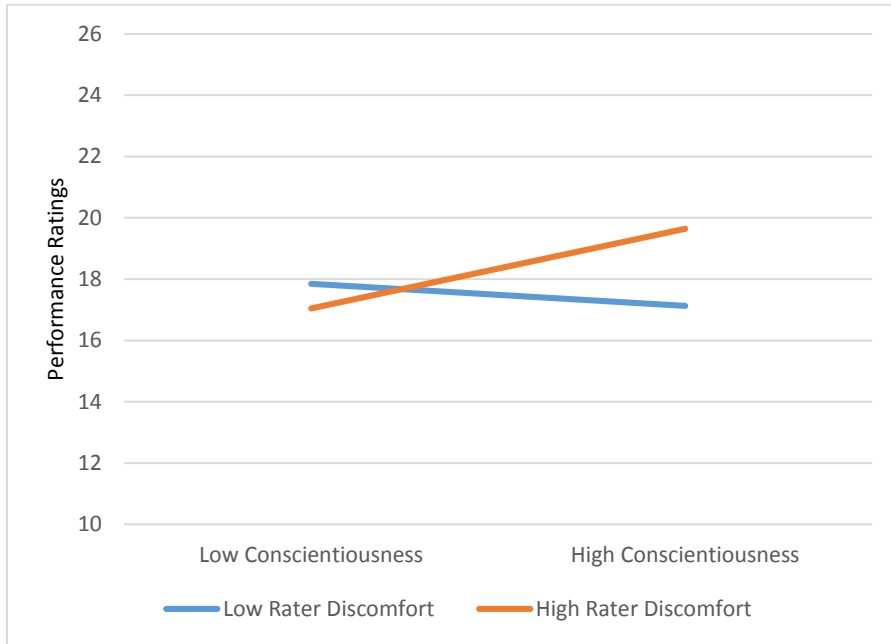


Figure 8. The moderational role of conscientiousness on the relationship between rater discomfort and performance ratings.

CHAPTER IV

Methods

In this chapter, the methodology of the present study is described. The chapter begins with a discussion of the participants and the power analysis used to determine the number of participants required for the current study. This is followed by the experimental design used to test the hypotheses listed at the end of Chapter 3. Next, an explanation of the instruments and procedures, including a description of tasks performed and research protocol, are outlined. The chapter concludes with a description of the data analysis plan.

Participants

To be eligible to participate, participants were required to be 18 years of age or older, provide an informed consent, have no knowledge of the study's hypotheses, and have no physical or cognitive handicap that would otherwise prohibit them from taking part in the study. A minimum sample size of 171 was required for the present study. This number was derived from an a priori power analysis conducted by the researcher, which is detailed later in this chapter. A non-probability convenience sampling technique was used as the sampling method in the present study, mainly because performance appraisals are largely universal. However, because various cultures conduct appraisals differently and practice different cultural norms and values, the generalization of this study's results are limited to the United States. All participants were required to be full-time employees with a minimum of one year performance appraisal experience. This demographic requirement was imposed so that measures of performance appraisal discomfort were

based on an individual's real-life experience as opposed to a guess about how an individual thinks he or she may feel about an appraisal experience. It is expected that the demographics of the sample will be representative of the greater population of managers who conduct performance appraisal. Moreover, the use of a convenience sample will allow inferences made from this study to generalize to greater populations.

Participants were recruited in two ways. One way was through the personal and professional network of the author. The second was through a survey panels service (Qualtrics) which assists researchers with collecting data by providing access to their partnership with market researchers for a fee. For this study, Clear Voice Research, a market research panel was used. Clear Voice Research utilizes a database of 12 million members, although for the purposes of this study, only individuals living in the United States were contacted. When recruited, members are told that by joining ClearVoiceSurveys.com opinion panel they would be invited to participate in online market research surveys in exchange for various incentives (See Appendix A). The sampling process employs simple randomization to give a representative sample of new and old members. All panelists were invited to participate in the current study via email invitation that included a link to the survey, the approximate length of the survey and the reward amount. To discourage 'professional survey takers' and survey fatigue, Clear Voice enforces a multi-panel membership policy and panelists are limited to one completed survey every 10 days.

A total of 180 (n=180) individuals participated in the current study, with 61.1% currently receiving performance appraisals and 16.7% not, while 22.2% were not

currently employed or the item was not applicable. Similarly, 45.6% of participants were currently employed in a job where they rated the performance of other employees, 32.2% did not and 22.2% were not currently employed or the item was not applicable. For gender, 41.3% of participants identified as male, 58.1% identified as female and 0.6% declined to state. For ethnicity, 74.4% identified as White, 6.7% identified as Black or African-American, 0.6% identified as American Indian or Alaska Native, 12.8% identified as Asian, 0.6% identified as Native Hawaiian or Other Pacific Islander, and 4.9% identified as Other. For age, the average age of participants was 28.34 with an SD of 13.07.

Participants had an average of 8.86 years of performance appraisal experience, with an SD of 9.24. For participants who conducted performance appraisals, an average of 6.43 employees is the number of employees they usually conducted appraisals for, with an SD of 9.93. Participants conducted performance appraisals an average of 2.66 times per year, with an SD of 1.73.

Protection of human participants. Through an informed consent form, all participants were informed that their participation was voluntary and that they may have refused to participate or may have ceased participation at any time during the study without assuming any consequences. Participants were told the purpose of the study and the time commitment required to complete the study activities. Participants were provided with contact information for both the researcher of this study and California School of Professional Psychology's Committee for the Protection of Human

Participants. All participant responses were anonymous and any potentially identifying information was kept confidential.

Power analysis. Among other uses, power analysis is used to calculate the required minimum sample size of a statistical analysis (Cohen, 1992). In addition to serving as a means of calculating an estimate of the minimum number of participants required, a power analysis also serves as an opportunity to review all aspects of an empirical study in order to increase the consistency of the methodology. Power represents the probability that a given statistical test will correctly reject the null hypothesis when the alternative hypothesis is true (Cohen, 1992). Although there are no formal standards for power, most researchers assess power for statistical tests at a value of .80 or .90 as a standard for adequacy (Cohen, 1992). Alpha represents the probability that a given statistical test will incorrectly reject the null hypothesis when the null hypothesis is true (Cohen, 1992). An alpha level of .05 is commonly used and is widely accepted as a standard of adequacy for testing one hypothesis (Cohen, 1992).

The components considered in the power analysis for the present study include: the number of tails (one-tailed or two-tailed), an effect size estimate for the main effect, an alpha (α error probability) value for the main effect, and a power ($1-\beta$ error probability) value (Cohen, 1992). The present study used a power analysis for the testing of all effects and interactions included in the six hypotheses. The effect sizes of all main effects were estimated, and a power analysis was calculated for the main effect with the most stringent r value, which is the smallest r value. The smaller the effect size, the more

stringent or larger the sample size needs to be. All other main effects did not need to be considered.

The estimate of required sample size for each hypothesis was calculated using G*Power3 (G*Power, 2009). For Hypothesis 1, the statistical test family of t test, means: difference between two independent means (two groups) was selected. Next, a priori was selected as the type of power analysis because it is used to calculate sample size given alpha, power, and estimated effect size, which are known. The alpha, power and effect size values were then entered into the fields generated by the selections described with $\alpha = .05$, power of .90, $d = .50$ and allocation ratio = 1. This yielded a minimum required sample size of 140. In addition to the main effect, there are several interactions being investigated in the current study. Specifically, the researcher is investigating: (a) the causal relationship between the level of performance and performance ratings (Hypothesis 1), (b) the moderating role of the level of performance on the relationship between rater discomfort and performance ratings (Hypothesis 2), (c) the mediating role of rater discomfort on the relationship between conscientiousness and performance ratings (Hypothesis 3), (d) the mediating role of rater discomfort on the relationship between agreeableness and performance ratings (Hypothesis 4), (e) the mediating role of rater discomfort on the relationship between self-efficacy in performance ratings and the level of performance rating (Hypothesis 5), and (f) the moderating role of conscientiousness on the relationship between rater discomfort and performance ratings (Hypothesis 6). Different types of power analyses were calculated for the interactions in this study and they are described below.

For Hypothesis 2, the statistical test family of F test, ANOVA: Fixed effects, special, main effects and interactions was selected. An a priori type of power analysis was chosen. The effect size, alpha, and power were then entered with $f = .25$, $\alpha = .05$, and a power of .90. This yielded a minimum required sample size of 171. The power analysis calculated for Hypothesis 2 yielded the most stringent results ($n=171$), which was used as the minimum requirement for number of participants for the present study. As this is a between subjects design, the 171 participants were randomly assigned to one of two groups, equaling in at least 85 participants per group.

For Hypothesis 3, F test was chosen for test family, Linear multiple regression: Fixed model, R^2 increase was chosen for statistical test, with a priori power analysis. The effect size, alpha, and power were then entered with $f^2 = .15$, $\alpha = .05$, power of .90 and number of predictors = 2. This yielded a minimum required sample size of 88.

For Hypothesis 4, F test was chosen for test family, Linear multiple regression: Fixed model, R^2 increase was chosen for statistical test, with a priori power analysis. The effect size, alpha, and power were then entered with $f^2 = .15$, $\alpha = .05$, power of .90 and number of predictors = 2. This yielded a minimum required sample size of 88.

As mentioned above, the power analysis calculated for Hypothesis 2 yielded the most stringent results ($n=171$), and was used as the minimum requirement for number of participants for the present study.

For Hypothesis 5, F test was chosen for test family, Linear multiple regression: Fixed model, R^2 increase was chosen for statistical test, with a priori power analysis. The

effect size, alpha, and power were then entered with $f^2 = .15$, $\alpha = .05$, power of .90 and number of predictors = 2. This yielded a minimum required sample size of 88.

For Hypothesis 6, the statistical test family of F test, ANOVA: Fixed effects, special, main effects and interactions was selected. An a priori type of power analysis was chosen. The effect size, alpha, and power were then entered with $f = .25$, $\alpha = .05$, and a power of .90. This yielded a minimum required sample size of 171. The power analysis calculated for Hypothesis 6 yielded equivalent results to those derived for Hypothesis 2, which was used as the minimum requirement for number of participants for the present study. As this is a between subjects design, the 171 participants were randomly assigned to one of two groups, equaling in at least 85 participants per group.

Design

The present study employed an experimental design with between- subjects variables. An experimental design was appropriate for the present study because it allowed the experiment to be controlled through a structured design, allowing the researcher to randomly assign participants to performance appraisal level conditions. Moreover, an experimental design shows increased internal validity as compared to a field study design (Mitchell & Jolley, 2001). The methodological limitations will further be discussed in the Discussion section.

Procedure

Prospective participants received an email from the researcher asking for their voluntary participation in the study. A brief description of the study and a URL link to the online survey was included in the email. The email stated that the purpose of the

study is to gain a better understanding of the decision-making processes that individuals undergo when they rate others' work performance. See Appendix B for the full written prompt sent to potential participants.

The survey began with a welcome page that reiterated the purpose of the study and required that participants view and agree to the terms of the informed consent. Per the recommendations of Podsakoff, MacKenzie, and Podsakoff (2012), the predictor variables were administered separately from the criterion variable and were conducted before the criterion variable was completed. Therefore, after consenting to the terms of the study, participants were asked to complete two questionnaires that measure: (a) conscientiousness and agreeableness, and (b) self-efficacy.

These two questionnaires were introduced in a counterbalanced order. Conscientiousness and agreeableness items were intermingled and randomized within the first questionnaire. The intermingling of the items from the two scales can be done because they use the same rating anchors and were developed as part of the same construct development process (Goldberg, 1999; Podsakoff et al., 2012).

After completing these questionnaires, participants were given a distraction task. The distraction task was administered before the manipulation of the independent variable with the purpose of acting as a temporal buffer between the measures of the predictor variables and the dependent measure (Podsakoff et al., 2012). The distraction task consisted of one scrambled phrase that participants needed to reenter, just as it appeared on their screen, using their computer keyboard. These scrambled phrases were

not real phrases nor did they have any meaning. Moreover, they usually contained slanted letters of various font types; an example includes Fidc7hW.

Following the distraction task, the level of performance was manipulated using a vignette. Participants were randomly assigned to either low performance or high performance conditions. This random assignment was executed through computer generated random assignment via the study's survey builder, Qualtrics. Participants were asked to read about a fictitious employee (See Appendix C). Following the manipulation, participants completed the performance appraisal discomfort scale. It is important to measure rater discomfort before performance ratings are measured, so that ratings have a chance to be a product of rater discomfort.

Following this, participants then rated the employee's performance, using the performance rating measure, based on the information that was provided in the vignettes. The performance rating measure (See Appendix D) consists of six components of employee performance as described in the next section.

Finally, participants were asked a number of demographic questions including age, gender, job title, and a series of questions pertaining to performance appraisal experience. The demographic variables were administered last so that considerations of appraisal experience did not distort the measurement of performance ratings and individual measures (Podsakoff et al., 2012). See Appendix E for the full text of the demographic questions. See Table 1 for a summary of the survey components and the order in which they will be administered.

Table 1

Procedure overview

Step	Measure
1*	Conscientiousness and Agreeableness domains of IPIP-NEO
2*	New General Self-efficacy scale (NGSE)
3	Distraction task
4	Manipulation: scenario A or B (random assignment)
5	Performance Appraisal Discomfort Scale (PADS)
6	Performance Rating (dependent measure)
7	Demographics Questionnaire

*Sequence of Steps 1 and 2 was counterbalanced

Instruments

The present study measured rater conscientiousness, rater agreeableness, rater discomfort, and self-efficacy. Rater conscientiousness and agreeableness were measured by the corresponding International Personality Item Pool Representation (IPIP-NEO) dimensions, rater discomfort by the Performance Appraisal Discomfort Scale (PADS), and self-efficacy by the New General Self-Efficacy Scale (NGSE).

The dependent variable of performance ratings was measured using a six-point employee performance review rating template.

Measure of conscientiousness and agreeableness: IPIP NEO. The NEO Personality Inventory was developed by Paul Costa & Robert McCrae in 1970 and has been revised several times to purge outdated and problematic items (McCrae, Costa & Martin, 2005). The original version of the measurement, published in 1978, was the Neuroticism-Extroversion-Openness Inventory. This version only measured three of the Big Five personality traits. It was later revised in 1985 to include all five traits and renamed the NEO Personality Inventory. Presently, there are revised versions of both the full and short version of the NEO, NEO-PI-3 and NEO-FFI-3, respectively. The original IPIP-NEO inventory contained 300 items. The newer, short version was designed to measure exactly the same traits as the original IPIP-NEO, but more efficiently with fewer items. The short version of the IPIP-NEO inventory uses 120 items from the original inventory. It is a systematic assessment of emotional, interpersonal, experimental, attitudinal, and motivational styles, for use in human resource development, industrial/organizational psychology, and vocational counseling & clinical practice

(McCrae & John, 1992). The NEO is a measure of the Five Factor Model of personality: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience.

At the European Conference on Personality in 1996, a new domain personality resource, the International Personality Item Pool (IPIP), was first introduced (Goldberg, 1999). The stimulus behind the IPIP was a perception that “the science of personality assessment has progressed at a dismally slow pace since the first personality inventories were developed over 75 years ago” (Goldberg, 1999, p. 7). In regard to personality-trait measurement, Goldberg (1999) attributed the seeming lack of progress in part to the policies and practices of commercial inventory publishers. For example, commercial publishers regularly require researchers to purchase and use an entire inventory as it is. In addition, some publishers of commercial inventories charge fees for researchers to access a scoring key. For these reasons, among others, Goldberg (1999) suggested that placing a set of personality items in the public domain might free researchers from the constraints imposed by copyrighted personality inventories. Hence, the International Personality Item Pool (IPIP) was created as an open-source personality inventory. The creator of the IPIP envisioned the IPIP website as “a computer-supported system that allows scientists to work with each other, facilities, and databases without regard to geographical location” (Finholt & Olson, 1997, p. 28). All of the IPIP items are correlated with the original inventory (NEO-PI-R), using a sample that has responded to both item pools. The coefficient alpha values, which represent the internal consistency of a scale, of the conscientiousness and agreeableness dimensions of the original NEO-PI-R, developed in

1985, are 0.90 and 0.86, respectively (Costa & McCrae, 2001). The inventory uses a five-point scale, ranging from strongly disagree to strongly agree. The current study used the conscientiousness and agreeableness dimensions of the short version of the IPIP inventory, each consisting of 24 items.

Measure of new general self-efficacy: NGSE. Self-efficacy is one's own competence to complete tasks and reach goals (Omrod, 2006) and has been defined by Bandura (1986) as an individual's judgment of his or her capability to organize and perform a course of action to attain some designated type of performance. The NGSE has been found to substantially contribute to organizational theory, research and practice (Chen et al., 2001). Also, the NGSE has been found to be unidimensional, have high reliability and have higher construct validity than the General Self-Efficacy Scale by Sherer (Chen et al., 2001). Moreover, the NGSE has demonstrated high reliability, predicted specific self-efficacy for a variety of tasks in various contexts and moderated the influence of previous performance on subsequent specific self-efficacy formation (Chen et al., 2001).

Measure of rating discomfort: PADS. The Performance Appraisal Discomfort Scale (PADS) is a micro-analytic application of the theory of job compatibility as described by Bernardin and Villanova (2005). Job compatibility refers to the extent to which employees maintain preferences for job characteristics that are consistent with the actual demands of the job. According to the job compatibility framework, employees whose preferences are at odds with actual job characteristics report greater discomfort in performing job activities and manifest behaviors indicative of less job involvement and

higher withdrawal and avoidance. Grounded in Bandura's theory of self-efficacy (Bandura, 1977), Abbott and Bernardin designed a 27-item questionnaire reflecting a variety of performance feedback situations (Bernardin & Villanova, 2005). In order to examine rater avoidance, researchers modified Abbott and Bernardin's (Cardy, Bernardin, Abbott, & Senderak, 1987) scale of self-efficacy for giving performance feedback (Villanova et al., 1993). The revised questionnaire, referred to as the Performance Appraisal Discomfort Scale (PADS), includes 20 of the original 27 items. Consistent with the original scale, the responses to the PADS reflect the degree of discomfort felt by raters in a variety of performance appraisal situations. Responses to PADS are provided using a five-point scale with anchors high discomfort to no discomfort. The mid-point response is undecided. A high score on this scale indicates a low degree of performance appraisal discomfort. The raters' level of discomfort in each type of situation reflects his or her feelings of self-efficacy when giving performance feedback. The reliability and validity of the PADS has been examined by several researchers since its creation. Villanova et al., (1993) reported alpha coefficients of .88 and .91. Another study reported a coefficient alpha for the PADS of .90 (Smith, Harrington & Houghton, 2000). Practical use of the scale has been recommended by Bernardin and Beatty (1984) where they suggest using discomfort scores to determine what training may be necessary for raters experiencing appraisal discomfort.

Measure of performance. This measure is a general short-form employee performance review. The review contains six dimensions including (1) job knowledge, (2) work quality, (3) attendance/punctuality, (4) initiative, (5) communication, and (6)

dependability. The review contains a five-point scale ranging from poor to excellent. The scores on these six dimensions are averaged to give an overall rating. This measure is entirely quantitative and will include no open-ended or qualitative questions. See Appendix D for a copy of the employee performance rating form.

CHAPTER V

Results

This chapter outlines the results of both the pilot study and the present study.

The Results of the Pilot Study

There were four questions in the pilot study:

1. How do you evaluate the overall performance of the employee?
2. How competent is the employee?
3. How qualified is the employee?
4. How likable is the employee?

The Pilot Study served as a manipulation check for the level of performance variable (See Appendix F). The results from the 20 respondents revealed that the low performing employee was found to be significantly less competent or qualified than the high performing employee. There was a significant difference in both the individual item scores and aggregate mean scores for low performance and high performance conditions; $t(18) = -13.17, p = 0.00$. Thus the manipulation of the scenarios was successful.

The secondary conclusion from the pilot study was that the conceptual midpoint on the 5-point Likert-type scale may not correspond with the value of 3. The conceptual midpoint may be 2.5. As such, the researcher considered either including another Likert point option in between fair and good or removing Likert-type point option very good, in order to balance the scale around the midpoint. In the end, we did not alter the rating scale of the measure for the current study because the items differed and the 5 point Likert-type scale more closely reflected performance appraisal scales used in applied

settings.

Descriptive Statistics

At the conclusion of data collection, the data analysis plan included a comprehensive data review to avoid errors in measurement associated with errors in data collection (Tabachnick & Fidell, 2007). The plan also called for the dataset to be proofread, potential outliers to be identified and addressed and missing data to be identified and addressed (Tabachnick & Fidell, 2007). Data management and data analysis was conducted using IBM SPSS Statistics 20 (IBM, 2011).

Table 2 provides the descriptive statistics for each variable in the present study, including minimum and maximum values, means, standard deviations, skewness and kurtosis. For the skewness and kurtosis values, the z score of either value was derived by dividing the statistic score by its standard error. Conscientiousness and self-efficacy were found to be negatively skewed. Self-efficacy is peaked and performance rating is flat. The Hartley F-max test yielded a value of 2.12 which shows that the different population groups showed similar variance.

Correlations among the five test variables can be found in Table 3.

Conscientiousness, Agreeableness, Self-efficacy, and Rater Discomfort have all been shown to correlate with one another. In contrast, performance ratings have not been shown to correlate with any of the other four measures. As seen in Table 3, people who are highly conscientious tend to be more agreeable, highly self-efficacious, and experience less rater discomfort. People who are highly agreeable tend to be highly self-efficacious and experience less rater discomfort. People who are highly self-efficacious

Table 2

Descriptive Statistics

Variable	Min	Max	Mean	SD	Skewness	Kurtosis
Conscientiousness	-11	40	21.57	12.54	-3.50	-1.37
Agreeableness	-11	40	16.09	10.88	-1.24	-1.70
Self-efficacy	19	56	46.19	7.65	-6.59	3.54
Rater Discomfort	20	94	45.04	16.21	2.57	-0.24
Performance Rating	6	30	17.69	7.87	0.19	-4.03

Table 3

Correlations among experimental measures

	Agreeableness	Self- efficacy	Rater Discomfort	Performance Rating
Conscientiousness	.597*	.581*	-.461*	.003
Agreeableness		.503*	-.336*	.023
Self-efficacy			-.375*	.101
Rater Discomfort				.061

Note: * indicates correlations that are significant at the .000 level

tend to experience less rater discomfort. However, none of the four variables have a significant correlation with performance rating.

Test of Assumptions

Prior to conducting the mediation analyses, the assumptions of linear regression and multiple regression were checked. The four principal assumptions which justify the use of linear regression models for purposes of prediction include (a) independence of cases, (b) normality, (c) linearity, and (d) homoscedasticity.

Independence of cases. Independence of cases is a function of research methodology as opposed to the other three principal assumptions which are measured through statistical means. Independence of cases is assured for all hypotheses due to the way all data was collected.

Normality. Normality was tested by the Kolmogorov-Smirnov test, by examining the skewness and kurtosis of the distribution, as well as identifying any potential outliers (Tabachnick & Fidell, 2007). As seen in Table 4, the Kolmogorov-Smirnov tests of normality revealed that all variables but agreeableness are not distributed normally. Visual inspection of the distribution graphs for each of the independent variables revealed relatively normal distributions with only slight variations from normal. The distribution for the dependent variable was shown to be somewhat bi-modal (See Appendix G), however this can be explained by the somewhat natural dichotomization of the data showing a high performer or low performer since those were the two conditions participants were given. Examination of the skewness and kurtosis showed that these values were mostly within the acceptable range.

Table 4

Kolmogorov-Smirnov Tests of Normality

	Statistic	df	Sig.
Performance Rating	.14	180	.00
Conscientiousness	.11	180	.00
Agreeableness	.07	180	.06
Self-efficacy	.18	180	.00
Rater Discomfort	.07	180	.02

Test of sub-group differences. To ensure that no group differences would affect the interpretation of hypothesis testing, the performance ratings assigned by candidates who were paid or unpaid were compared. Using a chi-square test, no differences were detected suggesting that both sub-groups performed the task similarly: $\chi^2(1, 24) = 18.213$, $p=0.793$.

Additionally, the performance ratings assigned by participants who held less than two years of performance appraisal experience were compared with participants who had two or more years of experience. It should be noted that no participants had less than one year of performance appraisal experience, as this was a requirement of the study. Using a chi-square test, no differences were detected suggesting that both sub-groups assigned ratings similarly: $\chi^2(1, 24) = 22.216$, $p=0.566$.

Linearity and homoscedasticity. Linearity and homoscedasticity were assessed by examining the scatter plots of residual values to determine whether the residuals have a straight-line relationship and that the variance is the same for all predicted scores (Tabachnick & Fidell, 2007). Upon inspection of a scatter plot of the residuals, all predictive relationships tested in the hypotheses were found to be linear and displayed homogeneity of variances. Levene's Test for Hypothesis 2 revealed that variance is equal across groups: $F(3, 176)=1.22$, $p=.30$. Levene's Test for Hypothesis 6 revealed that variance is equal across groups: $F(3,176) =2.20$, $p=.09$.

Tests of Hypotheses

Hypothesis 1. Hypothesis 1 states that when the level of performance indicates that actual performance is high, performance ratings will be higher. Hypothesis 1 was

tested using an independent samples T-test and assumptions of the statistical test were checked. The assumptions of T-test include the independence of cases, normality, and homogeneity of variance (Keppel & Wickens, 2004). The independence of cases is consistent with the research design. To support the independence of cases, the independent variable of the level of performance in Hypothesis 1 was dichotomous. Furthermore, a given rater's scores cannot influence the scores of other raters. As for normality, the definition of normal distribution is a continuous distribution in which the majority of data falls on or around the mean value, with about equal amount of variability on either end of the distribution (Tabachnick & Fidell, 2007). Normality was tested by examining the skewness and kurtosis of the distribution, as well as identifying any potential outliers (Tabachnick & Fidell, 2007). Homogeneity of variances was tested using Levene's Test for Equality of Variances to determine if the two conditions have about the same of different amounts of variability between scores.

Hypothesis 1 was supported. The mean of performance rating was significantly higher in the high actual performance condition ($M= 23.62$, $SD= 4.86$) than in the low actual performance condition ($M= 11.77$, $SD= 5.46$); $t(178)= 15.39$, $p= .000$ (See Figure 9). As seen in Table 5, this means that when presented with information that employee performance is high, the raters gave a higher performance rating.

Hypothesis 2. Hypothesis 2 states that there is a moderating effect of the level of performance on the relationship between rater discomfort and the level of performance rating. Hypothesis 2 was tested using a two-way analysis of variance (ANOVA) and

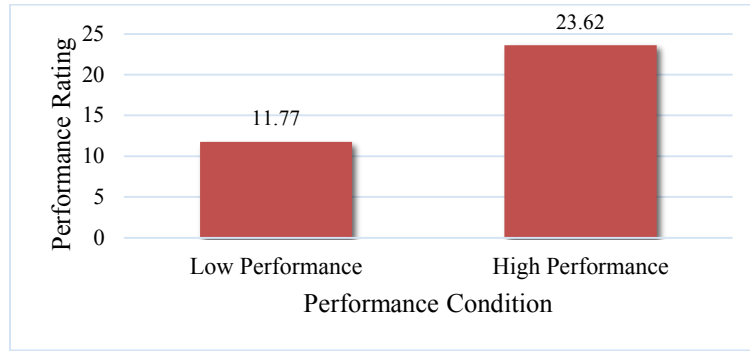


Figure 9. Bar graph displaying means for low and high performance conditions.

Table 5

Mean scores for low performance and high performance conditions

Condition	High		Low	
	Mean	SD	Mean	SD
Overall performance	4.20	0.79	1.30	0.48
Competency	4.20	0.63	1.50	0.53
Qualification	4.00	0.67	1.40	0.52
Likability	3.40	0.70	1.40	0.52
Total	3.95	0.75	1.40	0.50

appropriate post-hoc tests. Prior to conducting the ANOVA, the assumptions of the statistical test were checked. The assumptions of a two-way ANOVA are the independence of cases, normality, and homogeneity of variance (Keppel & Wickens, 2004). The independence of cases is consistent with the research design. The independent measures in Hypothesis 2 are continuous and were dichotomized using the median split method to generate two categories for each independent variable. Furthermore, a given rater's scores cannot influence the scores of other raters. A normal distribution is a continuous distribution in which the majority of data falls on or around the mean value, with about equal amount of variability on either end of the distribution (Tabachnick & Fidell, 2007). Normality was tested by examining the skewness and kurtosis of the distribution, as well as identifying any potential outliers (Tabachnick & Fidell, 2007). Homogeneity of variances was tested using Hartley's F-max test, which is conducted by calculating the ratio of the largest group variance as compared to the smallest group variance (Hartley, 1950).

Hypothesis 2 was supported. There was a significant moderational effect of the level of performance on the relationship between rater discomfort and the level of performance rating [$F(1, 176) = 4.97, p = .03$]. A small effect size for this moderating effect ($\eta^2 = .03$) was found. As seen in Figure 10, these results suggest that in the high performance condition, participants with high rater discomfort assigned lower ratings than participants with low rater discomfort. Interestingly, the findings for the high performance condition did not support the original hypothesis which stated that there would be no relationship between discomfort and ratings when performance was high. This supports the notion that people with discomfort tend to rate performance towards the middle of the rating scale. A post-hoc t-test revealed no

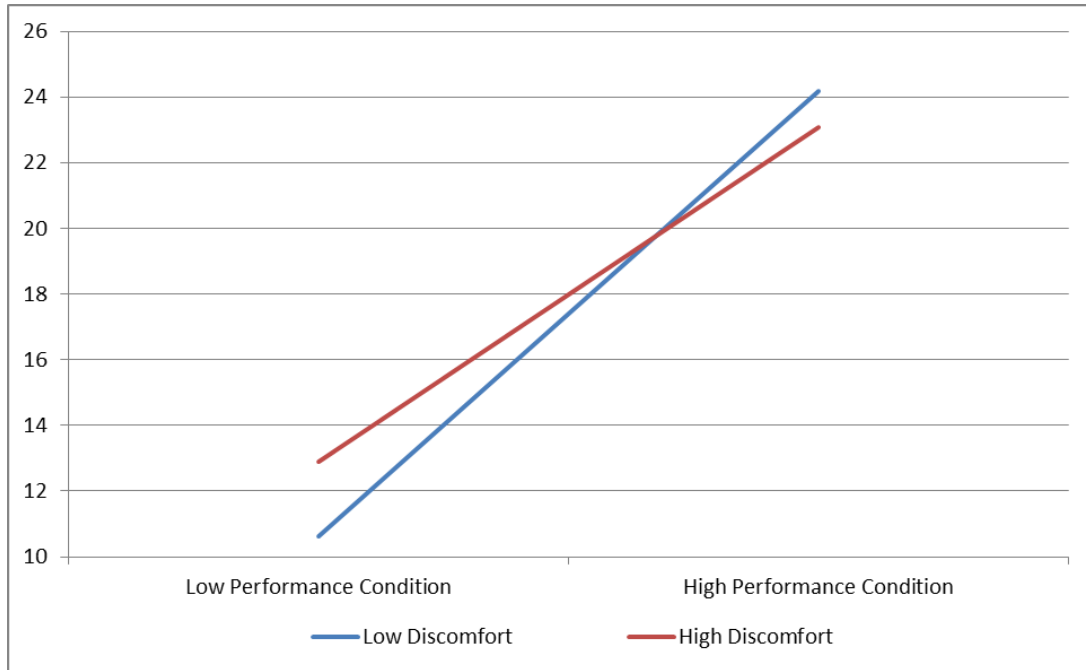


Figure 10. Mean performance ratings across levels of rater discomfort and information about performance.

significant difference in ratings assigned by low- and high-discomfort respondents within the high performance condition: $t(88)=-1.085, p=.281$. In the low performance condition, participants with high rater discomfort assigned higher ratings than participants with low rater discomfort. Additionally, the main effect of the level of performance on the level of performance rating was significant [$F(1,176) = 241.49, p=.00$], whereas the main effect of rater discomfort on the level of performance rating was not significant [$F(1,176) = .60, p=.44$]. As seen in Table 6, this indicated that the high performance rating condition ($M=23.62$) scored higher than the low performance rating condition ($M=11.77$).

Hypothesis 3. As seen in Figure 11, Hypothesis 3, states that rater discomfort mediates the relationship between conscientiousness and the level of performance ratings such that high levels of conscientiousness predict low levels of rater discomfort. Hypothesis 3 was tested using PROCESS, a macro developed by Hayes (2012) for use in PASW. This contemporary method of mediation analysis is preferred to Baron and Kenny's (1986) four-step method because it uses a path analysis framework to estimate the direct and indirect effects of an independent variable in a mediation model and implements bootstrapping, a statistical variance re-sampling technique that increases the stability of results (Hayes 2009; Preacher & Hayes, 2004). Bootstrapping is also helpful in the case that distribution-related assumptions are not met (Hayes 2009; Preacher & Hayes, 2004). Prior to conducting the mediation analyses, the assumptions of linear regression and multiple regression were checked. The assumptions of regression analysis include the absence of multicollinearity, normality, linearity, and homoscedasticity. A normal distribution is a continuous distribution in which the majority of data falls on or around the mean value, with about equal amount of variability on either end of the

Table 6

Mean levels of performance rating across performance conditions and rater discomfort groups

	High Performance Condition			Low Performance Condition			Total		
	Mean	SD	<i>n</i>	Mean	SD	<i>N</i>	Mean	SD	<i>n</i>
High Rater Discomfort	23.07	5.49	45	12.91	5.80	45	17.99	7.59	90
Low Rater Discomfort	24.18	4.13	45	10.62	4.90	45	17.40	8.17	90
Total	23.62	4.86	90	11.77	5.46	90	17.69	7.87	180

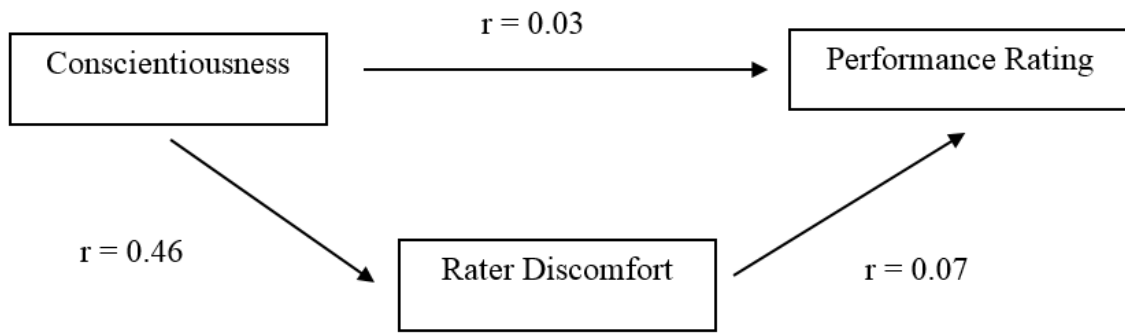


Figure 11. Rater discomfort's insignificant mediation on the relationship between conscientiousness and performance rating.

distribution (Tabachnick & Fidell, 2007). Normality was tested by examining the skewness and kurtosis of the distribution, as well as identifying any potential outliers (Tabachnick & Fidell, 2007). Linearity and homoscedasticity was assessed by examining the scatter plots of residual values to determine whether the residuals have a straight-line relationship and that the variance is the same for all predicted scores (Tabachnick & Fidell, 2007).

A Bootstrapping PROCESS Procedure was conducted to test this hypothesis. Rater discomfort was not found to mediate the relationship between conscientiousness and performance rating at a significant level ($\beta = -0.02$; LLCI = -0.07 ; ULCI = 0.03 ; $t(179) = 0.47$, $p = 0.63$). Thus, Hypothesis 3 was not supported, as can be seen in Table 7. The r scores in Figure 11 represent the direct effect coefficient, which shows the strength of the direct relationship between x , y , and m .

Hypothesis 4. As seen in Figure 12, Hypothesis 4 states that rater discomfort mediates the relationship between agreeableness and the level of performance rating. Hypotheses 4 was tested using Hayes' (2012) method of mediation analysis as outlined for Hypothesis 3. Specifically, a Bootstrapping PROCESS Procedure was conducted to test this hypothesis. Rater discomfort was not found to mediate the relationship between agreeableness and performance rating at a significant level ($\beta = -0.02$; LLCI = -0.07 ; ULCI = 0.02 ; $t(179) = 0.62$, $p = 0.54$). Thus, Hypothesis 4 was not supported, as seen in Table 8. The r scores in Figure 12, represent the direct effect coefficient, which shows the strength of the direct relationship between x , y , and m .

Hypothesis 5. As seen in Figure 13, Hypothesis 5, states that rater discomfort mediates the relationship between self-efficacy and the level of performance rating.

Table 7

Results of Bootstrapping Analysis for Hypothesis 3

	Indirect effect of x on y	Standard Error	t (p)	z (p)	95% Confidence Interval
Conscientiousness	-.02	.03	.47 (.64)	-.92 (.36)	-.07 to .03

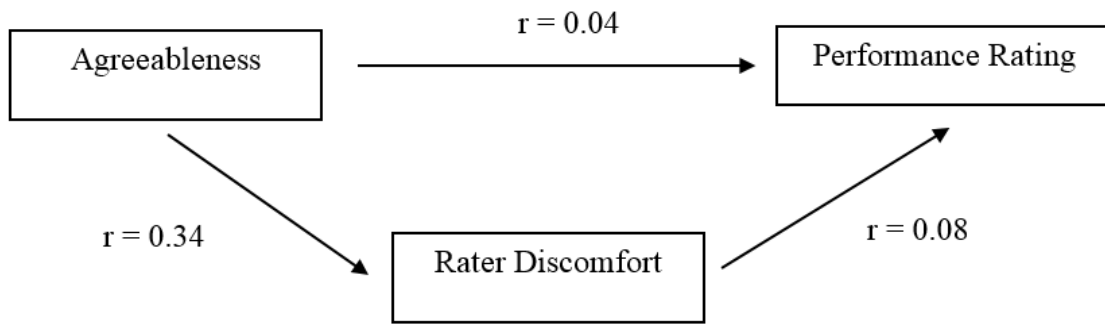


Figure 12. Rater discomfort’s insignificant mediation on the relationship between agreeableness and performance rating.

Table 8

Results of Bootstrapping Analysis for Hypothesis 4

	Indirect effect of x on y	Standard Error	t (p)	z (p)	95% Confidence Interval
Agreeableness	-.02	.02	.62 (.54)	-.93 (.35)	-.07 to .02

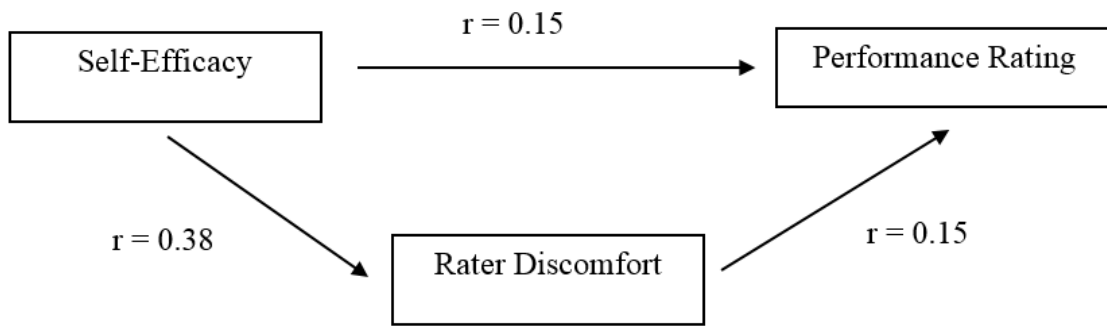


Figure 13. Rater discomfort mediating the relationship between self-efficacy and performance rating.

Hypotheses 5 was tested using Hayes' (2012) method of mediation analysis as outlined for Hypothesis 3. Specifically, a Bootstrapping PROCESS Procedure was conducted to test this hypothesis. Rater discomfort was not found to mediate the relationship between self-efficacy and performance rating at a significant level ($\beta = -0.04$; LLCI = -0.12 ; ULCI = 0.02 ; $t(179) = 1.80$, $p = 0.17$). Thus, Hypothesis 5 was not supported, as seen in Table 9. The r scores in Figure 13, represent the direct effect coefficient, which shows the strength of the direct relationship between x , y , and m .

Hypothesis 6. As seen in Figure 14, Hypothesis 6, states that there is a moderating effect of conscientiousness on the relationship between rater discomfort and the level of performance rating. Hypothesis 6 was not supported. A two-way between subjects ANOVA was conducted to test the hypothesis. There was no significant moderational effect of conscientiousness on the relationship between rater discomfort and the level of performance rating [$F(1, 176) = 1.88$, $p = .17$]. A small effect size for the moderating effect ($\eta^2 = .01$) was found. Neither the main effect of conscientiousness [$F(1, 176) = .60$, $p = .44$], nor the main effect of rater discomfort was significant [$F(1, 176) = .49$, $p = .49$]. Thus, Hypothesis 6 was not supported (See Table 10).

Hypothesis 6 was assessed using a two-way ANOVA and appropriate post-hoc tests, consistent with the manner described for Hypothesis 2.

Table 9

Results of Bootstrapping Analysis for Hypothesis 5

	Indirect effect of x on y	Standard Error	t (p)	z (p)	95% Confidence Interval
Self-efficacy	-.04	.04	1.80 (.07)	-1.36 (.17)	-.12 to .02

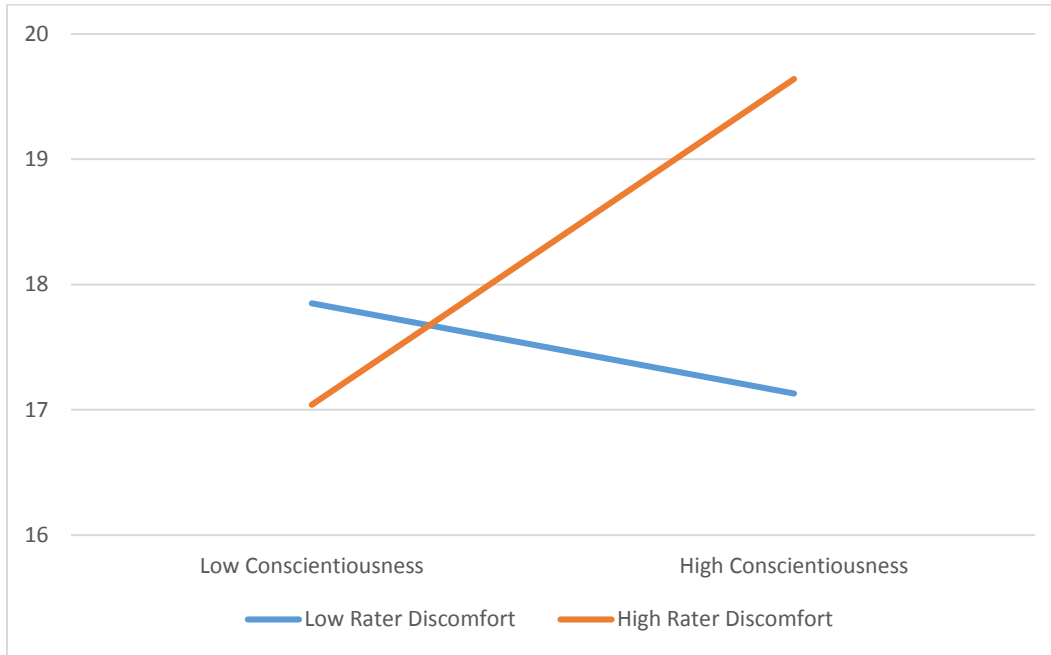


Figure 14. Mean performance ratings across levels of rater discomfort and conscientiousness.

Table 10

Mean levels of performance rating across conscientiousness and rater discomfort groups

	High Conscientiousness			Low Conscientiousness			Total		
	Mean	SD	<i>n</i>	Mean	SD	<i>n</i>	Mean	SD	<i>n</i>
High Rater Discomfort	19.64	7.39	33	17.04	7.60	57	17.99	7.59	90
Low Rater Discomfort	17.13	8.71	56	17.85	7.29	34	17.40	8.17	90
Total	18.06	8.30	91	17.34	7.46	89	17.69	7.87	180

CHAPTER VI

Discussion

In this final chapter, the present study is summarized and discussed, beginning with an interpretation of the findings. Next, the current findings are compared in relation to previous research. The limitations and considerations of the present study are then discussed. Lastly, the implications of the present study's findings are discussed followed by recommendations for future research.

Interpretation of Findings

The present study had several important findings, sequenced by hypothesis below.

Hypothesis 1. Supporting (Hypothesis 1) the main effect of the level of performance on performance rating was important. This means that the performance ratings assigned to high performers were higher than performance ratings assigned to low performers. This validates the establishment of the employee performance vignettes and performance rating scale that were developed and used in the study. This also emphasizes the importance of the level of performance information on the judgment of raters. This means that when shown performance level information, raters used the same pattern of factors to make judgments of performance and applied a rating that generally represented the specific performance level described. These findings contradict established research, which argues that in order for raters to rate accurately they need to be motivated to do so (Harris, 1994). The current study was absent of any of these motivations including situational factors like accountability; negative consequences, and rewards because they were not included. These findings show that even with absence of these motivational

factors, raters can still rate accurately. One can argue that a contrived environment such as the one created for the purposes of this study is not realistic, and that any real organizational setting would include motivations, either intrinsic or extrinsic. While this is a valid argument, it is interesting to see that when situational factors, negative consequences, and rewards are absent or not a factor, accurate performance ratings can be achieved. Ultimately, this implies that rater behavior may vary depending on the environment they are working in. As a result, raters may not require the above mentioned motivational factors to assign accurate ratings in all appraisal environments.

Hypothesis 2. The second most important finding of the present study is the moderating effect of the level of performance on the relationship between rater discomfort and performance rating (Hypothesis 2). This is particularly important because it provides support for compatibility theory, which gave birth to performance appraisal discomfort. Job compatibility theory asserts that the performance appraisal process is uncomfortable for many raters and influences ratings. More than that, this finding shows that depending on the type of performance information available (low or high performance), that discomfort may increase or decrease which in turn impacts performance ratings. In the low performance condition, those with high discomfort assigned higher ratings than those with low discomfort, as predicted. This finding aligns with the appraisal literature (Villanova et al., 1993) and suggests that leniency can be predicted by differences in ratee performance and rater discomfort, as measured by the PADS. These individuals may boost their ratings when performance is poor because they may be more concerned with what their peers and ratees may think of them based on their

assigned ratings. Furthermore, raters with low discomfort tended to assign more extreme ratings as compared to raters with high discomfort. Interestingly, in the high performance condition, raters with high discomfort assigned lower ratings than those with low discomfort. This finding, which did not support the study's hypothesis, does not align with the well-established theory that the discomfort associated with assigning a performance rating motivates individuals to inflate ratings in order to avoid feelings of discomfort (Villanova et al., 1993). In other words, even when performance was explicitly high, high discomfort individuals assigned lower ratings instead of inflating them. One explanation for this may be that uncomfortable raters engage in a thought process that focuses on striving to give a realistic rating for a high performing employee and what others may think of this assessment. This thought process can be taxing on a rater's mind and may devalue the importance of the appraisal task and compromise the validity of the performance rating as a result.

The remainder of the hypotheses in the study (Hypothesis 3, 4, 5, 6) were not supported, a review of these results are discussed below.

Hypothesis 3. Interestingly, rater discomfort did not mediate the relationship between the personality variable of conscientiousness, and performance ratings. Particularly for Hypothesis 3, the rationale was that conscientiousness raters are less willing to give in to the pressures associated with rater discomfort because the context of performance appraisal possesses specific features that are relevant to the trait of conscientiousness, such as accountability and the administrative significance of the ratings. The findings of this hypothesis did not align with much of the research literature

on rating leniency (Bernardin et al., 2000), which shows conscientiousness to be negatively correlated with rating level, so as conscientiousness increases, rating levels decrease. Previous research also suggests that rating behaviors are the product of relatively stable and reliable personality traits (Kane et al., 1995). It is important to note that this previous research focused on traits predicting performance ratings, whereas the current study examined how these traits influence rating inflation. The distinction here is that instead of stating that having particular traits predict an individual's ratings, the current study is examining how an individual's traits influence rating inflation under different performance conditions. Similar to the results of this study, Spence and Keeping (2009) also did not find conscientiousness to predict performance ratings. Including rater discomfort in the hypothesis built upon this established relationship within the theory and aimed to expand this area of research. However, because conscientious raters in this study were not found to produce less inflated ratings, further research is needed to examine the role of rater discomfort on conscientiousness and ratings before these results can be interpreted. Altering the methodology of the study in order to eliminate the effect the lack of buy-in from participants may produce different results and allow for a more confident interpretation of this finding. These limiting factors are further discussed in the limitations section.

Hypothesis 4. Specifically for Hypothesis 4, the rationale was that the act of performance appraisal can be a source of discomfort for those who are highly agreeable as compared to others because of the potential social repercussions associated with performance ratings and feedback (Villanova et al., 1993). Therefore, highly agreeable

individuals are likely to focus more on the relationship aspect rather than the task at hand and experience increased feelings of rater discomfort. The findings of this hypothesis did not align with relationship the rating leniency research literature (Bernardin et al., 2000). Similar to Hypothesis 3 above, although the majority of previous research suggests that rating behaviors are the product of relatively stable and reliable personality traits (Kane et al., 1995), there are studies that produced similar results to this study. For example, Spence and Keeping (2009) also did not find agreeableness to predict performance ratings. Including rater discomfort in the hypothesis built upon this established relationship and aimed to expand this area of research. However, because this hypothesis was not supported, further research is needed to explore the role rater discomfort plays with personality variables and performance ratings, similar to Hypothesis 3.

Hypothesis 5. Hypothesis 5, which hypothesized that rater discomfort mediates the relationship between self-efficacy and the level of performance ratings was also unsupported. Similar to Hypothesis 3, it was argued that being stronger in the traits of conscientiousness (used in Hypothesis 3) and self-efficacy provides an individual with more immunity to rater discomfort. The results of this hypothesis did not align with existing literature, which states that more self-efficacious raters are less influenced by rater discomfort when it comes to assigning performance ratings (Bernardin & Villanova, 2005). Additionally, there was no significance found between self-efficacy and performance ratings. This misalignment with the existing literature, opens the door to consider the context of these personality variables. Conscientiousness and self-efficacy are both traits that have importance when you are considering the task at hand. This is

usually when an individual is performing a task where they are actively participating, making judgments, and considering an infinite number of factors simultaneously, like how they are being perceived, performance, and the impact of their actions, for example. Perhaps the contrived setting of this study downgraded the importance of these variables such that they never had a chance to contribute as a legitimate factor because the situation did not allow for it. This issue is discussed further in the limitations section of the paper.

Hypothesis 6. Hypothesis 6 was also not confirmed, suggesting that conscientiousness does not moderate the relationship between rater discomfort and performance rating. A significant difference was not found between raters with high conscientiousness and low conscientiousness. Levels of conscientiousness were not shown to impact the relationship between discomfort and ratings. In other words, in a scenario when a rater has low or high discomfort, whether the person is a conscientious individual or not will not alter the impact discomfort may have on ratings. On the relationship between conscientiousness and performance ratings, the absence of this personality effect may be the result of the strong situation faced by participants. Personality theorists have long suggested that the effects of personality on behavior can be moderated by the strength of the situation (Cantor & Mischel, 1977; Weiss & Adler, 1984). Cantor and Mischel (1977) proposed that scenarios which lack ambiguity, labeled strong situations, can reduce the expression of personality. In the field of organizational psychology, researchers have found support for these claims. For example, researchers have found personality to be more predictive of specific workplace behaviors in scenarios that allow for a high degree of autonomy (Lee, Ashford, & Bobko, 1990; Simmering, Colquitt, Noe, & Porter, 2003), with autonomous situations being weaker than situations that do not allow for autonomy. The employee

vignettes that were shown to participants in this study were clear and concise, and likely created a very strong situation. Consequently, the study may have limited the opportunity for participants' dispositional tendencies (ex. Conscientiousness) to influence their performance ratings.

Relationship of Current Findings to Previous Research

The current study blends confirming the relationship of existing relationships with the performance appraisal literature and exploring new relationships based on various findings from other organizational psychology researchers. To the author's knowledge, while other studies have examined rater discomfort and performance ratings (Smith et al., 2000; Villanova et al., 1993) and individual differences and rating leniency (Bernardin et al., 2000; Spence & Keeping, 2009; Tziner et al., 2005), the present study is the only study that has examined the specific interactions between rater discomfort, the level of performance and performance ratings as well as individual differences, rater discomfort and performance ratings. Through this perspective, the present study can also be viewed as an extension of two lines of research: rater discomfort theory and individual differences and rater leniency.

First, the present study is an extension of research that distinguishes between the existence of rater discomfort and how it impacts performance ratings. Significance differences in performance rating found when presented with differing levels of performance information provide further evidence that the presence of rater discomfort, particularly when presented with specific ratee performance conditions is an important one. Whereas previously it was asserted that rating leniency was motivated by individual differences (Bernardin et al., 2000) and intentional bias (Harris, 1994), more recent

studies have provided evidence that rater discomfort does play a role in rating leniency (Smith et al., 2000).

Second, the present study is a continuation of research examining individual differences as determinants of rater leniency. The current study did not align with several other studies examining individual differences and rating leniency (Bernardin et al., 2000; Spence & Keeping, 2009; Tziner et al., 2005), where individual differences of conscientiousness and agreeableness were shown to affect performance ratings. This occurrence can be largely explained by the differing designs used in the current study as compared to other studies within the research literature. As discussed above, the current study used a contrived setting where paper people were used and raters may have lacked the needed buy-in required to examine a performance appraisal process and draw conclusions from it. Other studies have employed more realistic procedure, where the researchers had access to a real work setting. Some examples include, undergraduate students working in a group setting and evaluating each other's work (Bernardin et al., 2000) and first tour U.S. Army soldiers participating in a 180 degree performance appraisal process (Borman, White, & Dorsey, 1995). There have also been other studies which have employed a more experimental design to study performance appraisal. One study manipulated information about performance through vignettes which contained varying sales figures and situational constraints (Jawahar, 2005). Even with an experimental design, this study found individual differences to play a significant role and change the ratings that were assigned. Given this, it is still conceivable to use a quasi-experimental to study leniency in performance ratings and be able to detect differences

among assigned ratings. However, the current findings show that this method may not be ideal, especially when aiming to expand the appraisal literature.

Limitations and Considerations

The following section discusses potential limitations to the current study. The first limitation includes the use of an experimental design with relatively low external validity compared to a field study. This includes the study's manipulation of employee performance level instead of examining real world employee performance or actual performance data from a case study. Another component that lacked external validity included the study participants rating the performance of a fictitious employee. This limitation may have impacted the buy-in and realness factor for participants since there were no real stakes at hand. Although some of the participants in this study were given money in exchange for their participation, no differences were found between the group of participants who received payment and the group who did not. In a real world work setting, there are countless factors that are included when managers assess an employee's performance. These factors can be both variable and constant and include interpersonal relationship, organizational norms, and current workplace needs and attitudes, for example.

The online setting was another limitation of the study. There was no interaction between rater and ratee involved, besides reading the employee vignette. In actual organizations, there are typically added layers of continuous check-in and accountability regarding the performance appraisal process throughout a calendar year. This can involve managers conducting goal-setting and check-in meetings with employees at various

points of time throughout the year to discuss the entire performance ratings lifecycle, which includes goal setting, aligning accountabilities between managers and employees, managers rating performance every quarter with employees also tracking their progress, feedback sessions, and performance linked to bonus pay or incentives. It is important to note that the online setting of the study is not the sole limitation. In reality, many organizations employ a computer-based performance management system and use an online medium for employees and managers to set accountabilities and assign ratings. Asking a random participant to rate a fictitious written document, like the current study, is not a comparable substitute for conducting a performance appraisal in a work setting. If the study had employed the use of video, in addition to written documents that may have given more buy-in and realness factor to the task at hand. Using a video where participants could observe a fictitious employee performing job tasks would enrich the realism of the experiment, as opposed to rating the performance of a paper person. To increase the stakes, the study could have told participants that they were going to confront the ratee either virtually (e.g., video chat) or in-person, after they had assigned their ratings.

In addition to these factors, the present study makes no assessment or manipulation of many of other factors that have been shown to affect performance ratings. As discussed in the literature review, these factors include gender (Tsui & Gutek, 1984; Lee & Alvares, 1977; Lyness & Heilman, 2006; Maas & Gonzalez, 2011), ethnicity (Baron & Sackett, 2008), goal-setting (Tziner et al., 2001), motivation (Harris, 1994), time delay (Heneman & Wexley, 1983), organizational factors (Thomas, 1999)

and multi-rater feedback (Holzbach, 1978; Kerst, 2000). It is difficult for a study such as the current one to mimic this real life performance appraisal process, mainly because of all the variables, individuals and steps involved. Moreover, it is not advised or realistic for a research study to encompass all of these variables. With that said, without an explicit manipulation or assessment of all of these related factors, one cannot be completely certain that the findings of the present study are not due to the absence of these other factors.

After viewing the personality inventories, and performance vignette which employed paper people, the participants of this study may have seen through the experiment and responded in a biased way. For example, participants may have changed their performance ratings because the memory of responding to a self-efficacy or discomfort scale was salient and fresh in their minds. Murphy, Herr, Lockhart, and Maguire (1986) noted that paper people studies in which raters read performance vignettes and then rate performance of hypothetical rates are less realistic than are behavior observations studies where ratings are based on direct or indirect (e.g., video) observation of ratees' behavior. These researchers reported average effect sizes to be a bit larger in paper people studies ($d=.42$) than in behavior observation studies ($d=.31$). However, this difference was largely restricted to studies on the effects of variation in true performance level and the effects appraisal purpose. Since this finding, other researchers cite Murphy et al. (1986) when arguing against paper people studies. Although, one could argue that in many jobs (e.g., loan officer), outcome of behaviors (e.g., number of loans issued, dollar value of loans, etc.) are probably just as important as

observable behaviors in influencing performance evaluations (Jawahar, 2005). Thus, paper people studies may simulate important features of actual appraisals. In response to the results reported by Murphy et al., (1986) other researchers examined paper people studies more recently (Jawahar & Williams, 1997). These researchers contrasted four paper people studies with 16 behavior observation studies and found opposing results, effect sizes were larger in behavior observation studies than in paper people studies. Given such inconsistency, it may be premature to dismiss the usefulness of paper people design. This design allows for more control against hypothetical confounding variables and can be useful when the primary focus is on testing theoretical hypotheses.

No differences were found between the ratings provided when examining appraisal experience. Much of the research shows participants with more years of performance appraisal experience to provide lower performance ratings than those with less performance appraisal experience. The research on expertise shows that experts possess richer domain-specific knowledge structures and process information at a deeper level compared to novices (Chase & Ericsson, 1982; Chi et al., 1981). The thinking is that more experienced raters are not as susceptible to the pressures associated with assigning more lenient ratings. Overall, performance appraisal experience is not a very popular variable within appraisal research. Future research may focus on directly applying the research on expertise to more thoroughly examine the impact of appraisal experience on performance ratings.

The last limitation of the study was the use of the NGSE to measure self-efficacy. The use of a performance appraisal specific self-efficacy scale could have been used to

measure an individual's feelings about assigning a rating, specifically. Perhaps, employing this more task specific scale would have produced significant interaction between self-efficacy and performance ratings. Rater discomfort, as measured by the performance appraisal discomfort scale, has been shown to be a relatively stable rater characteristic that is not subject to significant change as a result of moderate experience in performance appraisal (Villanova et al., 1993).

Implications and Future Research

The current study contributes to the performance appraisal literature by demonstrating the extent to which discomfort can impact a rater's performance rating. The results demonstrate that leniency can be predicted by differences in rater performance and rater discomfort, as measured by the PADS. While the idea that raters are not fully objective when rating employees' performance is not a new one, there have been very few empirical investigations which studied the extent to which both rater attributions (discomfort, individual differences) and ratee performance can influence the accuracy of performance appraisals. Based on the findings of the current study, suggesting that performance appraisal discussions or actions are likely to produce discomfort, regardless of individual difference, organizations should consider the importance of rater discomfort and become educated on ways in which it can be less impactful in distorting ratings. One way organizations can combat rater discomfort in their supervisors is through training and coaching and the inclusion of goal-setting.

Recently, Murphy et al. (2004) found that different raters pursue different goals when rating performance and that these goals are stable across time. The current study

did not measure rater motivations, however as discussed above, the absence of motivation Studies including rater motivation will help researchers conceptualize the rating process as more than an evaluative exercise. By doing so, it is likely that we will get a more representative and, therefore, more accurate understanding of what happens during the rating process. An increased understanding of the performance appraisal process will enable practioners to improve the practice of performance appraisals. On a related note, the current study did not explicitly assign a role of peer or supervisor, however, there is cause to believe that peers may use different patterns of factors and cues when making judgments of performance, regardless of performance level. Results from a factor analysis showed that raters from different organizational levels use different factors in making ratings (Klimoski & London, 1974). Borman and Hallam (1991) found that supervisor and peer raters of clerical workers assigned different ratings when using behaviorally anchored rating scales. Further research examining various organizational levels could shed more light on this interesting finding.

Several recommendations are also noted here for future researchers. First, to maximize external validity, steps should be taken to study an environment where the performance that is being assessed is real or experienced somehow, as opposed to paper people. In addition, to maximize the generalizability of results, it is recommended that research of this nature be conducted in work settings or strictly by researchers with access to participants who are involved with performance appraisals. Moreover, it would be interesting to replicate the design of this study but instead of manipulating performance through performance vignettes, use real individual or group performance as some other

studies have done (Villanova et al., 1993). A discussion on the pros and cons of using paper people is in the previous section. It is also recommended for the larger performance appraisal literature to expand its conceptualization of performance appraisals and accept that accuracy may not be the primary goal of raters, as first proposed by Spence and Keeping, 2009. By adapting this way of thinking, we may begin to see a more representative and as a result, more accurate understanding of what happens during the rating process.

An increased understanding of the performance appraisal process will enable practitioners to improve the practice of performance appraisals. For example, performance appraisal trainings could be generated to address how raters consider non-performance factors. Current appraisal training, like frame-of-reference training (ex. Sulsky & Day, 1992) are designed to address cognitive errors. To the author's knowledge, no training exists to address intentional bias. Addressing the various factors that can influence ratings will help to make sure that trainings are aligned with the complexity and nuance of the performance appraisal process. Additionally, organizations can employ the use of a behaviorally-anchored rating scale (BARS). This is an appraisal method that aims to combine the benefits of narratives, critical incidents and quantified ratings by anchoring a quantified scale with specific narrative examples of good, moderate and poor performance (Schwab, Heneman, & DeCotiis, 1975). Behavior-based rating formats are generally superior to other formats in fostering performance improvement; when used with performance feedback, they tend to facilitate clarification

of work roles for employees and the reduction of role ambiguity and conflict (Tziner & Falbe, 1990).

In order to establish trainings, researchers will need to study rating influence and empirically establish the strength and direction of these influences, and then design trainings based on the empirical findings. Relatedly, organizations and raters may resist this training, since not all leniency may be bad in the eyes of a rater (e.g., personal gain). As a result, conceptualizing performance appraisals as an issue broader than the act of assigning a performance rating would assist to showcase rater behaviors as part of an integrated organizational process involving complex human interactions.

REFERENCES

- Antonioni, D. (1996). Designing an effective 360-degree appraisal feedback process. *Organizational Dynamics*, 25(2), 24-38. doi:10.1016/S0090-2616(96)90023-6
- Argyle, M., & Kendon, A. (1967). The experimental analysis of social performance. *Advances in Experimental Social Psychology*, 3, 55-98.
- Arthur, J. B. (1994). Effects of human resource systems on manufacturing performance and turnover. *The Academy of Management Journal*, 37(3), 670-687. doi:10.2307/256705
- Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology*, 49, 141-168. doi:10.1146/annurev.psych.49.1.141
- Atkins, P. B., & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology*, 55(4), 871-904. doi:10.1111/j.1744-6570.2002.tb00133.x
- Atwater, L. E., Waldman, D. A., & Brett, J. F. (2002). Understanding and optimizing multisource feedback. *Human Resource Management*, 41(2), 193-208. doi:10.1002/hrm.10031
- Atwater, L., Wang, M., Smither, J. W., & Fleenor, J. W. (2009). Are cultural characteristics associated with the relationship between self and others' ratings of leadership? *Journal of Applied Psychology*, 94(4), 876-886. doi:10.1037/a0014561
- Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology*, 77(6), 836-874. doi:10.1037/0021-9010.77.6.836
- Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology*, 77(6), 975-985. doi:10.1037/0021-9010.77.6.975
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 191-215. doi:10.1037/0033-295X.84.2.191
- Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology*, 4, 359-373. doi:10.1521/jscp.1986.4.3.359
- Bandura, A. (1988). Organizational Application of Social Cognitive Theory. *Australian Journal of Management*, 13(2), 275-302. doi:10.1177/031289628801300210

- Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology*, 38, 2, 335-345. doi:10.1111/j.1744-6570.1985.tb00551.x
- Barrett, R. S. (1966). *Performance rating*. Oxford, England: Science Research Assoc.
- Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1), 1-26. doi:10.1111/j.1744-6570.1991.tb00688.x
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173-1182. doi:10.1037/0022-3514.51.6.1173
- Bass, B. M. (1956). Reducing leniency in merit ratings. *Personnel Psychology*, 9(3), 359-369. doi:10.1111/j.1744-6570.1956.tb01074.x
- Becker, B. E., & Huselid, M. A. (1998). High performance work systems and firm performance: A synthesis of research and managerial implications. *Research in Personnel and Human Resources Management*, 16(1), 53-102. doi:10.1.1.319.7549
- Beer, M. (1978). A performance management system: Research, design, introduction and evaluation. *Personnel Psychology*, 31(3), 505-535. doi:10.1111/j.1744-6570.1978.tb00460.x
- Bernardin, H. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, 63(3), 301-308. doi:10.1037/0021-9010.63.3.301
- Bernardin, H., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work*. Boston, MA: Kent.
- Bernardin, H., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, 6, 205-212. doi: 10.5465/AMR.1981.4287782
- Bernardin, H., & Cardy, R. L. (1982). Appraisal accuracy: The ability and motivation to remember the past. *Public Personnel Management*, 11(4), 352-357. Retrieved from <http://psycnet.apa.org/psycinfo/1983-11458-001>
- Bernardin, H., Cooke, D. K., & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology*, 85(2), 232-234. doi:10.1037/0021-9010.85.2.232

- Bernardin, H., & Villanova, P. (1986). Performance appraisal. In Lock E. A. (Ed.) *Generalizing from laboratory to field settings* (pp. 43-62). Lexington, MA: Lexington.
- Bernardin, H., & Villanova, P. (2005). Research streams in rater self-efficacy. *Group & Organization Management, 30*(1), 61-68. doi:10.1177/1059601104267675
- Bernardin, H., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology, 62*(1), 64-69. doi:10.1037/0021-9010.62.1.64
- Bernthal, P., Sumlin, R., Davis, P., & Rogers, B. (1997). *Performance management practices survey report*. Pittsburgh, PA: Development Decisions International.
- Bettenhausen, K. L., & Fedor, D. B. (1997). Peer and upward appraisals: A comparison of their benefits and problems. *Group & Organization Management, 22*(2), 236-263. doi:10.1177/1059601197222006
- Bigoness, W, P. (1976). Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. *Journal of Applied Psychology, 61*(1), 80-84. doi:10.1037/0021-9010.61.1.80
- Billikopf, G. (2006). *Labor management in agriculture: Cultivating personnel productivity*, 2nd Edition. Berkeley, CA: University of California, Berkeley.
- Blum, M. L. & Naylor, J. C. (1968). *Industrial psychology: Its theoretical and social foundations*. New York, NY: Harper & Row.
- Block, J. (1993). Studying personality the long way. In D. Funder, R. Parke, C. Tomlinson-Keasy, & K. Widaman (Eds.), *Studying lives through time: Approaches to personality and development* (pp. 9-41). Washington, DC: American Psychological Association.
- Booz, E. G., Allen, J. L., & Hamilton C. L. (1982). *New product management for the 1980s*. New York, NY: Booz Allen Hamilton.
- Borman, W. C., & Hallam, G. L. (1991). Observation accuracy for assessors of work-sample performance: Consistency across task and individual-differences correlates. *Journal of Applied Psychology, 76*(1), 11-18. doi:10.1037/0021-9010.76.1.11
- Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology, 80*(1), 168-177. doi:10.1037/0021-9010.80.1.168

- Bracken, D.W., and Paul, K.B. (1993). The effects of scale type and demographics on upward feedback. In Society for Industrial and Organizational Society Annual Conference, May, San Francisco, CA. *Journal of Applied Psychology*, 86(5), 930-942. doi:10.1007/s10869-011-9218-5
- Brett, J. F., & Atwater, L. E. (2001). 360-degree feedback: Accuracy, reactions, and perceptions of usefulness. *Journal of Applied Psychology*, 86(5), 930-942. doi:10.1037/0021-9010.86.5.930
- Bretz, R. D., Milkovich, G. T. & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management*, 18(2), 321-352. doi:10.1177/014920639201800206
- Brinkerhoff, D. W., & Kanter, R. M. (1980). Appraising the performance of performance appraisal. *Sloan Management Review*, 21(3), 3-16. Retrieved from <http://europepmc.org/abstract/MED/10249849>
- Cable, D. M., & Judge, T. A. (1994). Pay preferences and job search decisions: A person-organization fit perspective. *Personnel Psychology*, 47(2), 317-348. doi:10.1111/j.1744-6570.1994.tb01727.x
- Campbell, J., & Pritchard, R. (1976). Motivation theory in industrial and organizational psychology. In M. D. Dunnette (Ed.) *Handbook of industrial and organizational psychology*. Chicago, IL: Rand-McNally.
- Cantor, N., & Mischel, W. (1977). Traits as prototypes: Effects on recognition memory. *Journal of Personality and Social Psychology*, 35(1), 38-48. doi:10.1037/0022-3514.35.1.38
- Cardy, R. L., Bernardin, H., Abbott, J. G., & Senderak, M. P. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational Psychology*, 60(3), 197-205. doi:10.1111/j.2044-8325.1987.tb00253.x
- Cardy, R. L., & Dobbins, G. H. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. *Journal of Applied Psychology*, 71(4), 672-678. doi:10.1037/0021-9010.71.4.672
- Cardy, R. L., & Kehoe, J. F. (1984). Rater selective attention ability and appraisal effectiveness: The effect of a cognitive style on the accuracy of differentiation among ratees. *Journal of Applied Psychology*, 69(4), 589-594. doi:10.1037/0021-9010.69.4.589
- Carroll, S. J. & Schneier, C. E. (1982). *Performance appraisal and review systems: The identification, measurement, and development of performance in organizations*. Glenview, Illinois: Scott, Foresman.

- Carson, K., Cardy, R., & Dobbins, G. (1991). Performance appraisal as effective management or deadly management disease: Two empirical investigations. *Group and Organizational Studies, 16*, 143-159. doi:10.1177/105960119101600203
- Cascio, W. F., & Valenzi, E. R. (1977). Behaviorally anchored rating scales: Effects of education and job experience of raters and ratees. *Journal of Applied Psychology, 62*(3), 278-282. doi:10.1037/0021-9010.62.3.278
- Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16, pp. 1-58). New York, NY: Academic Press.
- Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), 55-81. doi:10.1016/0010-0285(73)90004-2
- Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods, 4*(1), 62-83. doi:10.1177/109442810141004
- Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152. doi:10.1207/s15516709cog0502_2
- Chiles, A. M., & Zorn, T. E. (1995). Empowerment in organizations: Employees' perceptions of the influences on empowerment. *Journal of Applied Communication Research, 23*(1), 1-25. doi:10.1080/00909889509365411
- Claus, L., & Briscoe, D. (2009). Employee performance management across borders: A review of relevant academic literature. *International Journal of Management Reviews, 11*(2), 175-196. doi:10.1111/j.1468-2370.2008.00237.x
- Cleveland, J. N., & Murphy, K. R. (1992). Analyzing performance appraisals goal-directed behavior. In G. Ferris & K. Rowland (Eds.), *Research in personnel and human resources management* (Vol. 10, pp.121-185). Greenwich, CT: JAI Press.
- Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*(1), 130-135. doi:10.1037/0021-9010.74.1.130
- Coens, T., & Jenkins, M. (2002). *Abolishing performance appraisals: Why they backfire and what to do instead*. San Francisco, CA: Berrett-Koehler.
- Cohen, J (1992). A power primer. *Psychological Bulletin, 112* (1), 155-159. doi:10.1037/0033-2909.112.1.155

- Coladarci, T. (1992). Teachers' sense of efficacy and commitment to teaching. *The Journal of Experimental Education*, 60(4), 323-337.
doi:10.1080/00220973.1992.9943869
- Conn, S.R., & Rieke, M.L. (1994). *The 16PF fifth edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing.
- Cooper, W. H. (1981). Conceptual similarity as a source of illusory halo in job performance ratings. *Journal of Applied Psychology*, 66(3), 302-307.
doi:10.1037/0021-9010.66.3.302
- Costa, P.T., & McCrae, R.R. (1985). *The NEO personality inventory manual*. Odessa, FL: Psychological Assessment Resources.
- Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.
- Costa, P. T., & McCrae, R. R. (2011). The five-factor model, five-factor theory, and interpersonal psychology. In L. M. Horowitz, S. Strack (Eds.), *Handbook of interpersonal psychology: Theory, research, assessment, and therapeutic interventions* (pp. 91-104). Hoboken, NJ: John Wiley.
- Cronbach, L. (1955). Processes affecting scores on 'understanding of others' and assumed similarity. *Psychological Bulletin*, 52(3), 177-193.
doi:10.1037/h0044919
- Curtis, A. B., Harvey, R. D., & Ravden, D. (2005). Sources of political distortions in performance appraisals: Appraisal purpose and rater accountability. *Group & Organization Management*, 30(1), 42-60. doi:10.1177/1059601104267666
- Decotiis, T., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. *The Academy of Management Review*, 3(3), 635-646.
doi:10.5465/AMR.1978.4305904
- Delery, J. E., & Doty, D. H. (1996). Model of theorizing in strategic human resource management: Tests of universalistic, contingency, and configurational performance predictions. *The Academy of Management Journal*, 39(4), 802-835.
doi:10.2307/256713
- Delery, J. E., & Shaw, J. D. (2001). The strategic management of people in work organizations: Review, synthesis, and extension. In G. R. Ferris (Ed.), *Research in personnel and human resource management* (Vol. 20, pp. 165-197). Greenwich, CT: JAI Press.
- Deming, W. (1986). *Out of crisis*. Cambridge, MA: MIT Press.

- Demuijnck, G. (2009). Non-discrimination in human resource management as a moral obligation. *Journal of Business Ethics*, 88(1), 83-101. doi:10.1007/s10551-009-0100-6
- 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(6), 717-737. doi:10.1037/0021-9010.81.6.717
- DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, 93, 880-896. doi:10.1037/0022-3514.93.5.880
- Dipboye, R. L. (1985). Some neglected variables in research on discrimination in appraisals. *The Academy of Management Review*, 10(1), 116-127. doi:10.5465/AMR.1985.4277365
- Drenth, P. D. (1998). Personnel appraisal. In P. D. Drenth, H. Thierry, C. J. de Wolff (Eds.), *Handbook of work and organizational psychology (2nd ed.)*, Vol. 3: *Personnel psychology* (pp. 59-87). Hove, England: Taylor & Francis.
- Drucker, P. (1954). *The principles of management*. New York, NY: Harper Collins.
- Dulewicz, V. P. (1989). Performance appraisal and counseling. In P. Herriot (Ed), *Assessment and selection in organizations: Methods and practices for recruitment and appraisal* (pp. 645-649). New York, NY: John Wiley.
- Elmore, P. B., & LaPointe, K. A. (1975). Effect of teacher sex, student sex, and teacher warmth on the evaluation of college instructors. *Journal of Educational Psychology*, 67(3), 368-374. doi:10.1037/h0076608
- Eysenck, H. J. (1991). Dimensions of personality: The biosocial approach to personality. In J. Strelau, A. Angleitner (Eds.), *Explorations in temperament: International perspectives on theory and measurement* (pp. 87-103). New York, NY: Plenum Press.
- Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66(2), 127-148. doi:10.1037/0021-9010.66.2.127
- Finholt, T. A., & Olson, G. M. (1997). From laboratories to collaboratories: A new organizational form for scientific collaboration. *Psychological Science*, 8(1), 28-36. doi:10.1111/j.1467-9280.1997.tb00540.x
- Fleenor, J. W., & Prince, J. M. (1997). *Using 360-degree feedback in organizations: An annotated bibliography*. Greensboro, NC: Center for Creative Leadership.

- Fletcher, C. (2001). Performance appraisal and management: The developing research agenda. *Journal of Occupational and Organizational Psychology*, 74(4), 473-487. doi:10.1348/096317901167488
- Folger, R., Konovsky, M. A., & Cropanzano, R. (1992). A due process metaphor for performance appraisal. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 14, pp. 129-177). Greenwich, CT: JAI Press.
- Frayne, C. A., & Latham, G. P. (1987). Application of social learning theory to employee self-management of attendance. *Journal of Applied Psychology*, 72(3), 387-392. doi:10.1037/0021-9010.72.3.387
- Fried, Y., Tiegs, R. B., & Bellamy, A. R. (1992). Personal and interpersonal predictors of supervisors' avoidance of evaluating subordinates. *Journal of Applied Psychology*, 77(4), 462-468. doi:10.1037/0021-9010.77.4.462
- Gardner, D. G., & Pierce, J. L. (1998). Self-esteem and self-efficacy within the organizational context: An empirical examination. *Group and Organization Management*, 23, 48-70. doi:10.1177/1059601198231004
- Gaugler, B. B., & Rudolph, A. S. (1992). The influence of assesse performance variation on assessors' judgments. *Personnel Psychology*, 45(1), 77-98. doi:10.1111/j.1744-6570.1992.tb00845.x
- Geake, A., Oliver, K., & Farrell, C. (1998). *A survey of the views of HR practitioners on 360 degree processes*, Thames Ditton, England: SHL.
- Gerhart, B., & Milkovich, G. T. (1990). Organizational differences in managerial compensation and financial performance. *The Academy of Management Journal*, 33(4), 663-691. doi:10.2307/256286
- Gist, M. E., & Mitchell, R. R. (1992). Self-efficacy: A theoretical analysis of its determinants and malleability. *The Academy of Management Review*, 17(2), 183-211. doi:10.2307/258770
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48(1), 26-34. doi:10.1037/0003-066X.48.1.26
- Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, 7, 7-28. Retrieved from http://projects.ori.org/lrg/PDFs_papers/A%20broad-bandwidth%20inventory.pdf
- Gomez-Mejia, L. R., & Balkin, D. B. (1992). Determinants of faculty pay: An agency theory perspective. *The Academy of Management Journal*, 35(5), 921-955. doi:10.2307/256535

- Guion, R. M. & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology*, 18(2), 135-164. doi:10.1111/j.1744-6570.1965.tb00273.x
- Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, 16(1), 250-279. doi:10.1177/001872678603901104
- Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. (1974). Race and sex as determinants of ratings by potential employees in a simulated work shaping task. *Journal of Applied Psychology*, 59, 705-711. doi:10.1037/h0037503
- Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management*, 20(4), 737-756. doi:10.1016/0149-2063(94)90028-0
- Härtel, C. E. (1993). Rating format research revisited: Format effectiveness and acceptability depend on rater characteristics. *Journal of Applied Psychology*, 78(2), 212-217. doi:10.1037/0021-9010.78.2.212
- Hartley, H.O. (1950). The use of range in analysis of variance. *Biometrika*, 37, 271-280.
- Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76, 408-420. doi:10.1080/03637750903310360
- Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from <http://www.afhayes.com/public/process2012.pdf>
- Hauenstein, N. M. (1992). An information-processing approach to leniency in performance judgments. *Journal of Applied Psychology*, 77(4), 485-493. doi:10.1037/0021-9010.77.4.485
- Heneman, R. L., & Wexley, K. N. (1983). The effects of time delay in rating and amount of information observed on performance rating accuracy. *Academy of Management Journal*, 26(4), 677-686. doi:10.2307/255915
- Huber, V., Neale, M., & Northcraft, G. (1987). Decision bias and personnel selection strategies. *Organizational Behavior and Human Decision Processes*, 40, 136-147. doi:10.1016/0749-5978(87)90009-4
- Huselid, M. A. (1995). The impact of human resource management practices on turnover, productivity, and corporate financial performance. *The Academy of Management Journal*, 38(3), 635-672. doi:10.2307/256741

- Huselid, M. A., Jackson, S. E., & Schuler, R. S. (1997). Technical and strategic human resource management effectiveness as determinants of firm performance. *The Academy of Management Journal*, 40(1), 171-188. doi:10.2307/257025
- IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.
- Ichniowski, C., Levine, D. I., Olson, C., & Strauss, G. (2000). *The American workplace: Skills, compensation, and employee involvement*. New York, NY: Cambridge University Press.
- Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. *Research in Organizational Behavior*, 5, 141-197. Retrieved from <http://psycnet.apa.org/psycinfo/1984-10878-001>
- Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64(1), 349-371. doi:10.1037/0021-9010.64.4.349
- Ilgen, D. R., Mitchell, T. R., & Fredrickson, J. W. (1981). Poor performers: Supervisors' and subordinates' responses. *Organizational Behavior and Human Performance*, 27(3), 386-410. Retrieved from <http://www.sciencedirect.com/science/article/pii/0030507381900301>
- Jacobson, M. B., & Effertz, J. (1974). Sex roles and leadership: Perceptions of the leaders and the led. *Organizational and Human Performance*, 12(3), 383-396. doi:10.1016/0030-5073(74)90059-2
- Jensen-Campbell, L.A., Knack, J.M., Waldrip, A.M., & Campbell, S.D. (2007). Do Big Five personality traits associated with self-control influence the regulation of anger and aggression? *Journal of Research in Personality*, 41, 403-424. doi:10.1016/j.jrp.2006.05.001
- Judge, T. A., Erez, A., & Bono, J. E. (1998). The power of being positive: The relation between positive self-concept and job performance. *Human Performance*, 11(2-3), 167-187. doi:10.1207/s15327043hup1102&3_4
- Judge, T. A., Locke, E. A., Durham, C. C., & Kluger, A. N. (1998). Dispositional effects on job and life satisfaction: The role of core evaluations. *Journal of Applied Psychology*, 83(1), 17-34. doi:10.1037/0021-9010.83.1.17
- Joiner, B. L., & Scholtes, P. R. (1988). *Total quality leadership vs. management by control*. Center for Quality and Productivity Improvement, University of Wisconsin, Madison.

- Jones, E. E., Rock, L., Shaver, K. G., Goethals, G. R., & Ward, L. M. (1968). Pattern of performance and ability attribution: An unexpected primacy effect. *Journal of Personality and Social Psychology, 10*(4), 317-340. doi:10.1037/h0026818
- Jurgensen, C. E. (1950). Overall job success as a basis for employee ratings. *Journal of Applied Psychology, 34*(5), 333-337. doi:10.1037/h0054680
- Kane, J. S., Bernardin, H., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *The Academy of Management Journal, 38*(4), 1036-1051. doi:10.2307/256619
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461. doi:10.2307/1170678
- Kassin, S. M. (2005). On the psychology of confessions: Does innocence put innocents at risk? *American Psychologist, 60*(3), 215-228. doi:10.1037/0003-066X.60.3.215
- Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook*. Upper Saddle River, NJ: Prentice-Hall.
- Klimoski, R. J., & London, M. (1974). Role of the rater in performance appraisal. *Journal of Applied Psychology, 59*(4), 445-451. doi:10.1037/h0037332
- Klores, M. S. (1966). Rater bias in forced-distribution performance ratings. *Personnel Psychology, 19*(4), 411-421. doi:10.1111/j.1744-6570.1966.tb00315.x
- Kohn, A. (1999). *Punished by rewards: The trouble with gold stars, incentive plans, A's, and other bribes*. New York, NY: Houghton Mifflin Harcourt.
- Kozlowski, S. W., Chao, G. T., Morrison, R. F. (1998). Games raters play: Politics, strategies, and impression management in performance appraisal. In Smither, W. (Ed.), *Performance appraisal: State-of-the-art in practice* (pp. 163-205). San Francisco, CA: Jossey-Bass.
- Kramar, R., McGraw, P., & Schuler, R. S. (1997). *Human resource management in Australia*. Melbourne, AU: Longman.
- Kreitner, R., & Kinicki, A. (2007). *Organizational Behavior, 7th ed*, Avenues of the Americas, NY: McGraw Hill.
- Kubo, I., & Saka, A. (2002). An inquiry into the motivations of knowledge workers in the Japanese financial industry. *Journal of Knowledge Management, 6*(3), 262-271. doi:10.1108/13673270210434368

- Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology, 79*(3), 332-340. doi:10.1037/0021-9010.79.3.332
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107. doi:10.1037/0033-2909.87.1.72
- Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. New York, NY: Academic Press.
- Landy, F. J., Farr, J. L., Saal, F. E., & Freytag, W. R. (1976). Behaviorally anchored scales for rating the performance of police officers. *Journal of Applied Psychology, 61*(6), 750-758. doi:10.1037/0021-9010.61.1.750
- Landy, F. J., & Trumbo, D. A. (1976). *Psychology of work behavior*. Oxford, England: Dorsey.
- Larson, J. R. (1984). The performance feedback process: A preliminary model. *Organizational Behavior & Human Performance, 33*(1), 42-76. doi:10.1016/0030-5073(84)90011-4
- Latham, G. P., & Locke, E. A. (2007). New developments in and directions for goal-setting research. *European Psychologist, 12*(4), 290-300. doi:10.1027/1016-9040.12.4.290
- Latham, G. P., & Wexley, K. N. (1993). *Increasing productivity through performance appraisal*. Reading, MA: Addison-Wesley.
- Law, D. R. (2007). Appraising performance appraisals: A critical look at an external control management technique. *International Journal of Reality Therapy, 26*(2), 18-25. Retrieved from <http://0-web.a.ebscohost.com.library.alliant.edu/ehost/pdfviewer/pdfviewer?sid=035b531a-6be1-4725-8eaa-3cf2280db3d2%40sessionmgr4005&vid=104&hid=4112>
- Lawler, E. E. (1976). Participation and pay. *International Journal of Production Research, 14*(3), 367-372. doi:10.1080/00207547608956609
- Lawler, E. E. (1994). From job-based to competency-based organizations. *Journal of Organizational Behavior, 15*(1), 3-15. doi:10.1080/00207547608956609
- Lee, A. H., Chen, W. C., & Chang, C. J. (2008). A fuzzy AHP and BSC approach for evaluating performance of IT department in the manufacturing industry in Taiwan. *Expert Systems with Applications, 34*(1), 96-107. Retrieved from <http://dx.doi.org/10.1016/j.eswa.2006.08.022>

- Leslie, J. B., Gryskiewicz, N. D., & Dalton, M. A. (1998). Understanding cultural influences on 360-degree feedback process. In W. W. Tornow and M. London (Eds.), *Maximizing the value of 360-degree feedback: A process for successful individual and organizational development* (pp. 196-216). San Francisco, CA: Jossey-Bass.
- Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management*, 30(6), 881-905. doi:10.1016/j.jm.2004.06.005
- Likert, R. (1959). *Motivational approach to management development*. Cambridge, MA: Harvard Business Review.
- London, M., & Poplawski, J. R. (1976). Effects of information on stereotype development in performance appraisal and interview contexts. *Journal of Applied Psychology*, 61(2), 199-205. doi:10.1037/0021-9010.61.2.199
- London, M. & Smither, J. W. (2006). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance-related outcomes? Theory-based applications and directions for research. *Personnel Psychology*, 48(4), 803-839. doi:10.1111/j.1744-6570.1995.tb01782.x
- Longenecker, C. O., Sims, H. P., & Gioia, D. A. (1987), Behind the mask: The politics of employee appraisal. *The Academy of Management Executive*, 1(3), 183-193. doi:10.5465/AME.1987.4275731
- Mandell, M. M. (1956). Supervisory characteristics and ratings: A summary of recent research. *Personnel*, 32, 435-440. Retrieved from <http://psycnet.apa.org/psycinfo/1957-03873-001>
- MacDuffie J. P. (1995). Human resource bundles and manufacturing performance: Organizational logic and flexible production systems in the world auto industry. *Industrial and Labor Relations Review*, 48(2), 197-221. Retrieved from <http://www.jstor.org/discover/10.2307/2524483>
- Markle, G. L. (2000). *Catalytic coaching: The end of the performance review*. Portsmouth, NH: Quorum Books.
- Martocchio, J. J., & Judge, T. A. (1997). Relationship between conscientiousness and learning in employee training: Mediating influences of self-deception and self-efficacy. *Journal of Applied Psychology*, 82(5), 764-773. doi:10.1037/0021-9010.82.5.764
- Maurer, T. J., & Alexander, R. A. (1991). Contrast effects in behavioral measurement: An investigation of alternative process explanations. *Journal of Applied Psychology*, 76(1), 3-10. doi:10.1037/0021-9010.76.1.3

- McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1), 81-90. doi:10.1037/0022-3514.52.1.81
- McCrae, R. R., Costa, P. T., & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO personality inventory. *Journal of Personality Assessment*, 84(3), 261-270. doi:10.1207/s15327752jpa8403_05
- McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2), 175-215. doi:10.1111/j.1467-6494.1992.tb00970.x
- McGregor, D. (1957). *An uneasy look at performance appraisal*. Soldiers Field: Harvard Business Review.
- McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, 69(1), 147-156. doi:10.1037/0021-9010.69.1.147
- McNamara, C. (2005). *Field guide to consulting and organizational development: A collaborative and systems approach to performance, change and learning*. Minneapolis, MN: Authenticity Consulting.
- Meier, B. P., Robinson, M. D., & Wilkowski, B. M. (2006). Turning the other cheek: Agreeableness and the regulation of aggression-related primes. *Psychological Science*, 17, 136-142. doi:10.1111/j.1467-9280.2006.01676.x
- Mignerey, J. T., Rubin, R. B., & Gorden, W. I. (1995). An investigation of newcomer communication behavior and uncertainty. *Communication Research*, 22(1), 54-85. doi:10.1177/009365095022001003
- Milkovich, G. T., Wigdor, A. K. (1991). *Pay for performance: Evaluating performance appraisal and merit pay*. Washington, DC: National Academy Press.
- Mitchell, M., & Jolley, J. (2001). *Research Design Explained* (4th Edition) New York: Harcourt.
- Mohrman, S., & Lawler, E. (1983). *Performance appraisal revisited*. Los Angeles, CA: University of Southern California Center for Effective Organizations.
- Morrison, E. W. (1993). Newcomer information seeking: Exploring types, modes, sources, and outcomes. *Academy of Management Journal*, 36(3), 557-589. doi:10.2307/256592

- Morrison, E. W., Chen, Y., & Salgado, S. R. (2004). Cultural differences in newcomer feedback seeking: A comparison of the United States and Hong Kong. *Journal of Applied Psychology, 53*(1), 1-22. doi:10.1111/j.1464-0597.2004.00158.x
- Muchinsky, P. M. (2012). *Psychology applied to work*. Summerfield, NC: Hypergraphic Press.
- 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*(1), 39-44. doi:10.1037/0021-9010.71.1.39
- Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*(4), 619-624. doi:10.1037/0021-9010.74.4.619
- Murphy, K. R., & Cleveland, J. (1991). *Performance appraisal: An organizational perspective*. Boston: Allyn and Bacon.
- Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.
- Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology, 89*(1), 458-164. doi:10.1037/0021-9010.89.1.158
- Murphy, K. R., & Jako, R. (1989). Under what conditions are observed intercorrelations greater or smaller than true intercorrelations? *Journal of Applied Psychology, 74*(5), 827-830. doi:10.1037/0021-9010.74.5.827
- Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*(2), 218-225. doi:10.1037/0021-9010.78.2.218
- Napier, N. K., & Latham, G. P. (1986). Outcome expectancies of people who conduct performance appraisals. *Personnel Psychology, 39*(4), 827-837. doi:10.1111/j.1744-6570.1986.tb00597.x
- Nathan, B. R., & Alexander, R. A. (1985). The role of inferential accuracy in performance rating. *The Academy of Management Review, 10*(1), 109-115. doi:10.5465/AMR.1985.4277361
- Nathan, B. R., & Lord, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. *Journal of Applied Psychology, 68*(1), 102-114. doi:10.1037/0021-9010.68.1.102

- Neck, C. P., Stewart, G. L., & Manz, C. C. (1995). Thought self-leadership as a framework for enhancing the performance of performance appraisers. *The Journal of Applied Behavioral Science*, 31(3), 278-302. doi:10.1177/0021886395313004
- Neely, A., Richards., H., Mills, J., Platts, K., & Bourne, M. (1997). Designing performance measures: A structured approach. *International Journal of Operations and Production Management*, 17(11), 1131-1152. doi:10.1108/01443579710177888
- Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. New York, NY: Henry Holt.
- Ngo, H., Foley, S., Wong, A., & Loi, R. (2003). Who gets more of the pie?: Predictors of perceived gender inequity at work. *Journal of Business Ethics*, 45(3), 227-241. doi:10.1023/A:1024179524538
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231-259. doi:10.1037/0033-295X.84.3.231
- Omrod, J. E. (2006). *Educational Psychology*. (5th Edition) Upper Saddle River, NJ: Prentice Hall.
- Parker, J. W., Taylor, E. K., Barrett, R. S., & Martens, L. (1958). Rating scale content: Relationship between supervisory- and self-ratings. *Personnel Psychology*, 12(1), 49-63. doi:10.1111/j.1744-6570.1959.tb00796.x
- Peters, T. J., & Waterman, R. H. (1982). *In Search of Excellence: Lessons from America's best-run companies*. New York, NY: Harper Business Essentials.
- Pfau, B., Kay, I., & Nowack, K. M. (2002). Does 360-degree feedback negatively affect company performance? *HR Magazine*, 47, 55-59. Retrieved from http://219.151.4.130/guochen2/jixiaoguanli/contents/thesis/the_014_01.pdf
- Pollack, D. M., & Pollack, L. J. (1996). Using 360 degree feedback in performance appraisal. *Public Personnel Management*, 25, 507-528. doi:10.1177/009102609602500410
- Podsakoff, P. M., MacKenzie, S. B., Paine, J. B., & Bachrach, D. G. (2000). Organizational citizenship behaviors: A critical review of the theoretical and empirical literature and suggestions for future research. *Journal of Management*, 26(3), 513-563. doi:10.1177/014920630002600307
- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 63, 539-569. doi:10.1146/annurev-psych-120710-100452

- Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*, 36, 717-731. doi:10.3758/BF03206553
- Pulakos, E. D., Schmitt, N., & Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within ratees to measure halo. *Journal of Applied Psychology*, 71(1), 29-32. doi:10.1037/0021-9010.71.1.29
- Rosen, B., & Jerdee, T. H. (1974). Effects of applicant's sex and difficulty of job on evaluations of candidates for managerial positions. *Journal of Applied Psychology*, 59(4), 511-512. doi:10.1037/h0037323
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68-78. doi:10.1037/0003-066X.55.1.68
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413-428. doi:10.1037/0033-2909.88.2.413
- Saks, A. M. (1995). Longitudinal field investigation of the moderating and mediating effects of self-efficacy on the relationship between training and newcomer adjustment. *Journal of Applied Psychology*, 80(2), 211-225. doi:10.1037/0021-9010.80.2.211
- Salgado, J. F. (1998). Big five personality dimensions and job performance in army and civil occupations: A European perspective. *Human Performance*, 11(2-3), 271-288. doi:10.1207/s15327043hup1102&3_8
- Schneider, R. J., Hough, L. M., & Dunnette, M. D. (1996). Broadsided by broad traits: How to sink science in five dimensions or less. *Journal of Organizational Behavior*, 17(6), 639-655. doi:10.1002/(SICI)1099-1379(199611)17:6<639::AID-JOB3828>3.0.CO;2-9
- Scholtes, P. R. (1998). *The leader's handbook: Making things happen, getting things done*. New York, NY: McGraw-Hill.
- Schraeder, M., Becton, J. B., & Portis, R. (2007). A critical examination of performance appraisals, an organization's friend or foe? *The Journal for Quality & Participation*, 30(1), 20-25. Retrieved from <https://secure.asq.org/perl/msg.pl?prvurl=http://asq.org/quality-participation/2007/03/human-resources/critical-examination-performance-appraisals.pdf>
- Schraeder, M., & Jordan, M. (2011). Managing performance: A practical perspective on managing employee performance. *The Journal for Quality and Participation*,

34(2), 4-10. Retrieved from <https://secure.asq.org/perl/msg.pl?prvurl=/quality-participation/2011/07/human-resources/managing-performance-a-practical-perspective-on-managing-employee-performance.pdf>

Schwab, D. P., Heneman, H. G., & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology*, 28(4), 549-562. doi:10.1111/j.1744-6570.1975.tb01392.x

Scullen, S. E., Mount, M. K., Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85(6), 956-970. doi:10.1037/0021-9010.85.6.956

Sherer, M., & Adams, C. H. (1983). Construct validation of the self-efficacy scale. *Psychological Reports*, 53, 899-902. doi:10.2466/pr0.1983.53.3.899

Shipper, F., Hoffman, R. C., IV, & Rotondo, D. M. (2007). Does the 360 feedback process create actionable knowledge equally across cultures? *Academy of Management Learning & Education*, 6(1), 33– 35. doi:10.5465/AMLE.2007.24401701

Simmering, M. J. Colquitt, J. A. Noe, R. A., Porter, C. O. (2003). Conscientiousness, autonomy fit, and development: A longitudinal study. *Journal of Applied Psychology*, 88(5), 954-963. doi:10.1037/0021-9010.88.5.954

Smith, W. J., Harrington, K. V., & Houghton, J. D. (2000). Predictors of performance appraisal discomfort: A preliminary examination. *Public Personnel Management*, 29(1), 21-32. doi:10.1177/009102600002900102

Smither, J. W., London, M., & Richmond, K. (2005). The relationship between leaders' personality and their reactions to and use of multisource feedback: A longitudinal study. *Group & Organization Management*, 30(2), 181-210. doi:10.1177/1059601103254912

Solomonson, A. L., & Lance, C. E. (1997). Examination of the relationship between true halo and halo error in performance ratings. *Journal of Applied Psychology*, 82(5), 665-674. doi:10.1037/0021-9010.82.5.665

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, 36(11), 1202-1212. doi:10.1037/0022-3514.36.11.1202

Spector, P. E. (2003). Review of 'taking the measure of work: A guide to validated scales for organizational research and diagnosis'. *Personnel Psychology*, 56(3), 813-816. Retrieved from <http://0-web.a.ebscohost.com.library.alliant.edu/ehost/detail?vid=126&sid=035b531a-6be1-4725-8eaa-3cf2280db3d2%40sessionmgr4005&>

hid=4112&bdata=JnNpdGU9ZWhvc3QtbGl2ZSZzY29wZT1zaXRl#db=psyh&AN=2003-08215-027



- Spence, J. R., & Keeping, L. M. (2009). The impact of non-performance information on ratings of job performance: A policy-capturing approach. *Journal of Organizational Behavior, 31*, 587-608. doi:10.1002/job.648
- Spinks, N., Wells, B., & Meche, M. (1999). Appraising the appraisals: Computerized performance appraisal systems. *The Career Development International, 4*(2), 94-100. doi:10.1108/13620439910254713
- Stajkovic, A. D., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin, 124*(2), 240-261. doi:10.1037/0033-2909.124.2.240
- Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology, 77*(4), 501-510. doi:10.1037/0021-9010.77.4.501
- Tabachnick, B. G., Fidell, L. S. (2007). *Using multivariate statistics*. (5th Edition). Boston, MA: Pearson Education.
- Terpstra, D. E., & Rozell, E. J. (1993). The relationship of staffing practices to organizational level measures of performance. *Personnel Psychology, 46*(1), 27-48. doi:10.1111/j.1744-6570.1993.tb00866.x
- Tett, R. P., Jackson, D. N., Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*(4), 703-742. doi:10.1111/j.1744-6570.1991.tb00696.x
- Tsui, A. S., & Barry, B. (1986). Interpersonal affect and rating errors. *The Academy of Management Journal, 29*(3), 586-599. doi:10.2307/256225
- Tziner, A. (1999). The relationship between distal and proximal factors and the use of political considerations in performance appraisal. *Journal of Business and Psychology, 14*(1), 217-231. doi:10.1023/A:1022931106379
- Tziner, A., Murphy, K. R., & Cleveland, J. N. (2001). Relationships between attitudes toward organizations and performance appraisal systems and rating behavior. *International Journal of Selection and Assessment, 9*(3), 226-239. doi:10.1111/1468-2389.00176
- Tziner, A., & Falbe, C. M. (1990). Actual and preferred climates of achievement orientation and their congruency: An investigation of their relationships to work attitudes and performance in two occupational strata. *Journal of Organizational Behavior, 11*(2), 159-167. doi:10.1002/job.4030110207

- Tziner, A., & Murphy, K. R. (1999). Additional evidence of attitudinal influences in performance appraisal. *Journal of Business and Psychology, 13*(3), 407-419. doi:10.1023/A:1022982501606
- Tziner, A., Murphy, K. R., & Cleveland, J. N. (2005). Contextual and rater factors affecting rating behavior. *Group and Organization Management, 30*(1), 89-98. doi:10.1177/1059601104267920
- United States Office of Personnel Management. (2011). *A handbook for measuring employee performance: Aligning employee performance plans with organizational goals*. Washington, DC: US OPM.
- Usilaner, B., & Leitch, J. (1989). Miles to go...or unity at last. *Journal for Quality and Participation, 12*(2), 60-67. Retrieved from https://secure.asq.org/perl/msg.pl?prvurl=http://asq.org/data/subscriptions/jqp_open/1989/june/jqp12i2usilaner.pdf
- Villanova, P., & Bernardin, H. (1989). Impression management in the context of performance appraisal. In R. A. Giacalone, P. Rosenfeld (Eds.), *Impression management in the organization* (pp. 299-313). Hillsdale, NJ: Lawrence Erlbaum.
- Villanova, P., & Bernardin, H. (1991). Performance appraisal: The means, motive, and opportunity to manage impressions. In R. A. Giacalone, P. Rosenfeld (Eds.), *Applied impression management: How image-making affects managerial decisions* (pp. 81-96). Thousand Oaks, CA: Sage.
- Villanova, P., Bernardin, H., Dahmus, S. A., & Sims, R. L. (1993). Rater leniency and performance appraisal discomfort. *Educational and Psychological Measurement, 53*(3), 789-799. doi:10.1177/0013164493053003023
- Wanberg, C. R., & Kammeyer-Mueller, J. D. (2000). Predictors and outcomes of proactivity in the socialization process. *Journal of Applied Psychology, 85*(3), 373-385. doi:10.1037/0021-9010.85.3.373
- Wexley, K. N., Sanders, R. E., & Yukel, G. A. (1973). Training interviewers to eliminate contrast effects in employment interviews. *Journal of Applied Psychology, 57*(3), 233-236. doi:10.1037/h0034714
- Wherry, R. J., & Bartlett, C. J. (2006). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35*(3), 521-551. doi:10.1111/j.1744-6570.1982.tb02208.x
- Whisler, T. L. (1958). Performance appraisal and the organization man. *The Journal of Business, 31*(1), 19-27. Retrieved from <http://www.jstor.org/stable/2350957>

- Wang, X. M., Wong, K., & Kwong, J. Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology, 95*(3), 546-561. doi:10.1037/a0018866
- Weiss, H. M., & Adler, S. (1984). Personality and organizational behavior. *Research in Organizational Behavior, 6*, 1-50. Retrieved from <http://psycnet.apa.org/psycinfo/1984-30230-001>
- Wong, K., & Kwong, J. Y. (2007). Effects of rater goals on rating patterns: Evidence from an experimental field study. *Journal of Applied Psychology, 92*(2), 577-585. doi:10.1037/0021-9010.92.2.577
- Wood, R., & Bandura, A. (1989). Impact of conceptions of ability on self-regulatory mechanisms and complex decision making. *Journal of Personality and Social Psychology, 56*(3), 407-415. doi:10.1037/0022-3514.56.3.407
- Wright, J. C., & Mischel, W. (1987). A conditional approach to dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology, 53*(6), 1159-1177. doi:10.1037/0022-3514.53.6.1159
- Youndt, M. A., Snell, S. A., Dean, J. W., & Lepak, D. P. (1996). Human resource management, manufacturing strategy, and firm performance. *The Academy of Management Journal, 39*(4), 836-866. doi:10.2307/256714
- Yun, G. J., Donahue, L. M., Dudley, N. M., & McFarland, L. A. (2005). Rater personality, rating format, and social context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment, 13*(2), 97-107. doi:10.1111/j.0965-075X.2005.00304.x

APPENDIX A



Sample Qualtrics Survey Panel Invitation



Survey Invitation

Dear Reni,

You have been invited to participate in a Consumer Services Survey!

 Reward: \$1.50
 Survey Length: 15 minutes
▶ Click to Participate

Click the following link below if the button above does not work:

[Click Here to participate in this survey](#)

Thank you for your continued participation!

Sincerely,
The Entire Clear Voice Surveys Team
[Contact Us](#)

If you prefer to no longer hear from Clear Voice Surveys, please [unsubscribe](#). During the removal period, you may receive some invitations that were already in process when your request was received. To unsubscribe via U.S. Mail please send all requests to: Clear Voice Surveys, 1675 Larimer Street, Suite 640, Denver CO 80202 USA

APPENDIX B

Written invitation sent to potential participants

I would like to invite you to participate in an online survey that will take approximately 20-30 minutes of your time. I am conducting this study in partial fulfillment of my doctorate degree in Industrial/Organizational Psychology at California School of Professional Psychology. The purpose of the study is to gain a better understanding of the decision-making processes that individuals undergo when they rate others' work performance. Your contributions will be completely anonymous and no identifying information will be collected. Your participation is voluntary and much appreciated.

If you have any questions about this survey or would like to know the results when the study has concluded, please contact Mina Azizi, mazizi@alliant.edu.

[Survey Link]

Thank you,

Mina Azizi

APPENDIX C

Employee Performance Vignettes

High Performance:

S.H. has worked for Clover Corporation for nearly three years in the same role. During this time, S.H. has been able to help train and mentor new hires. Tasks and projects completed by S.H. have occasionally been identified by management as appropriate for use as templates for others' projects. S.H. provides notice when occasionally unable to attend office meetings and contributes actively to office meetings when present. S.H. has minimal absences due to unexpected events such as illness and submits requests for paid time off in advance. S.H. identifies opportunities for process improvements and solutions to issues encountered by staff. After project completion, S.H. routinely confirms with clients that all requirements of a project have been met. S.H. takes notes when assigned new tasks and enters new tasks into the department work schedule. S.H. collaborates with other team members when developing a work schedule to ensure all deadline are met.

Low Performance:

C.T. has worked for Clover Corporation for nearly three years in the same role. During this time, C.T. has often required assistance from others when completing more complicated tasks. Review of tasks and projects completed by C.T. have occasionally been found to have errors and require revisions. C.T. sometimes submits request for time off without the company-required 5-day notice. About once a month, C.T. is five to fifteen minutes late for office meetings due to having lost track of time while performing

other work tasks. C.T. contributes to office meetings by identifying issues encountered by staff. C.T. has occasionally shown poor follow-through on assigned tasks, failing to follow up with clients after project completion. C.T. sometimes needs to be reminded of tasks assigned and does not make consistent use of the department work schedule. Especially when performing important or complex tasks, C.T. does not always take advantage of the email and conference call channels of communication to keep other team members updated with pertinent information.

APPENDIX D

Employee Performance Rating Form

Instructions: Please rate the employee described in the scenario across the six categories of job performance using the scale provided.

Poor Fair Good Very Good Excellent

Job Knowledge

Work Quality

Attendance/Punctuality

Initiative

Communication

Dependability

APPENDIX E

Demographics

1. How many years of performance appraisal experience do you have?
2. In your current employment, do you receive appraisal on your job performance?
3. Are you currently employed in a job where you rate the performance of other employees?
4. For how many employees, if any, do you usually conduct performance appraisals?
5. How many employees report to you?
6. How many employees have you rated on a performance appraisal in your career so far?
7. How many times per calendar year does your organization require you to participate in performance appraisals?
8. What is your age?
9. Gender
10. Your current job title

APPENDIX F

Pilot Study

To depict a low or high performing employee, general job performance information based on the six dimensions of the performance rating were included, instead of direct statements that may lead to demand characteristics. These dimensions included job knowledge, communication, attendance, dependability, initiative, and work quality. An example includes “Review of tasks and projects completed by C.T. have occasionally been found to have errors and require revisions”. Both vignettes were designed to be approximately the same length in order to prevent participants from drawing any obvious conclusions about the importance of one case. In order to control the variation in both conditions; no description or demographic information was included in either vignette. Statements were developed by considering common work behavior and practices that could be categorized as low performance and high performance. Subtle cues indicating frequency (ex. occasionally, routinely) were used when constructing statements to suggest a high or low performing condition. Past studies have manipulated performance in different ways. For example, an objective performance cue, such as overt performance was included to provide explicit information about how well an employee performed (Spence & Keeping, 2009). In the same study, performance was manipulated by overtly stating that the employee had either below average, average or above average performance (Spence & Keeping, 2009). A high level of performance example includes, over the past year, James has been an above average performer (Spence & Keeping, 2009). Similar to the current study, absolute statements were not used to minimize

possible differences in interpretation. Another study used peer evaluation scores and a 15 minute videotape from *The Apprentice* to demonstrate or manipulate high and low performance (Wang et al., 2010). One risk in using peer evaluations is the difficulty for raters to ignore between-individual comparisons when giving absolute ratings (Wong & Kwong, 2005). In another study, company standards of performance for various job types were created by the authors (Lee, Welbourne, Hoke, & Beggs, 2008). For example, “disc jockeys need to facilitate, on average, five promotional giveaways per month. They are expected to show up on time for each work shift” (Lee et al., 2008, p. 452). Then, descriptions depicting good or poor performance were juxtaposed with those company performance standards. For example, “Over the past year, worker has averaged six prize giveaways per month. Worker has been reliable in showing up to work on time” (Lee et al., 2008, p. 452).

The between subjects design feature of the experiment prevented any potential carryover effects. After reading the vignette, participants in the pilot study were given the following instructions, and then asked to answer the following questions in regards to the employee depicted in the vignette on a five-point Likert-type scale:

Instruction: Please answer each question below regarding the employee described in the scenario. Please rate the employee described in the scenario using the scale provided.

The pilot study used the same performance rating scale as the present study, but did not include the scale categories because they were not relevant to the questions in the pilot study. This scale ranged from poor (1) to excellent (5).

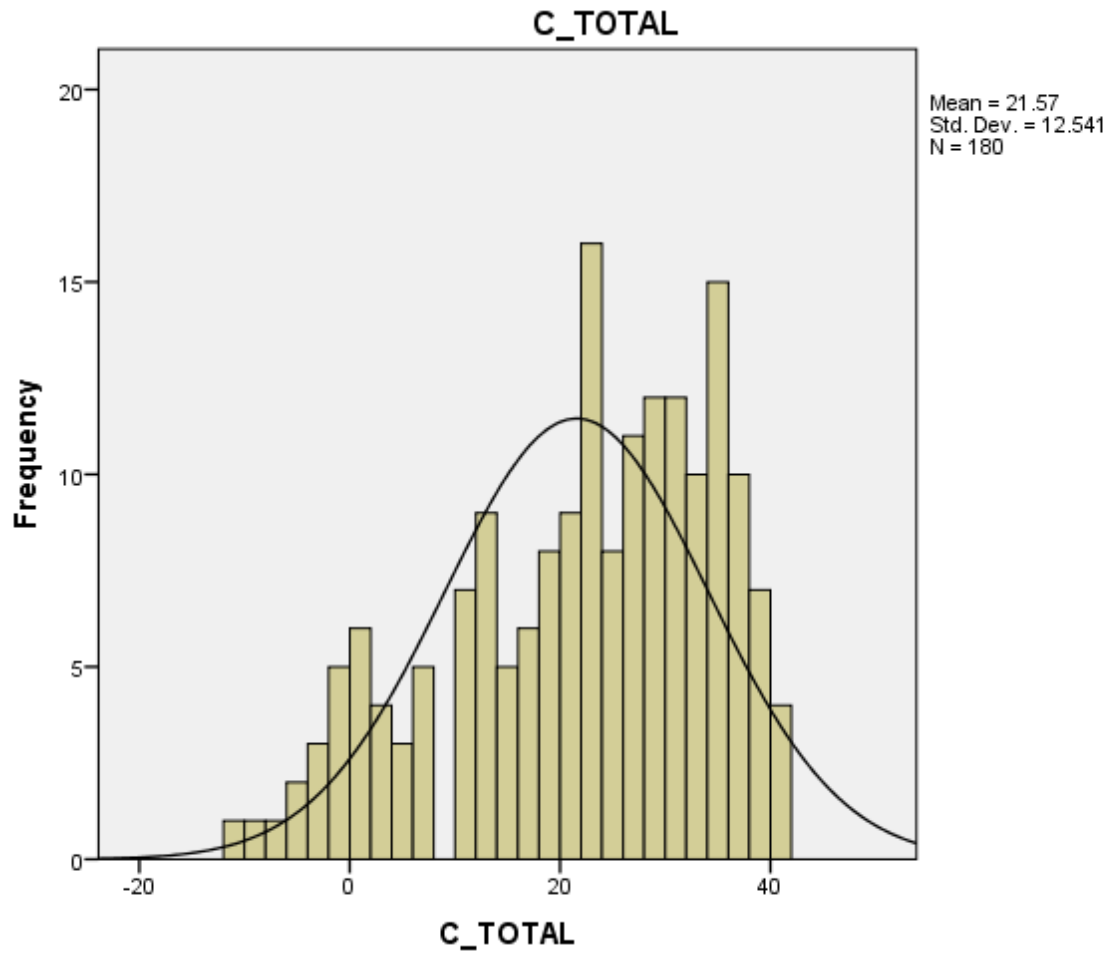
1. How do you evaluate the overall performance of the employee?

2. How competent is the employee?
3. How qualified is the employee?
4. How likable is the employee?

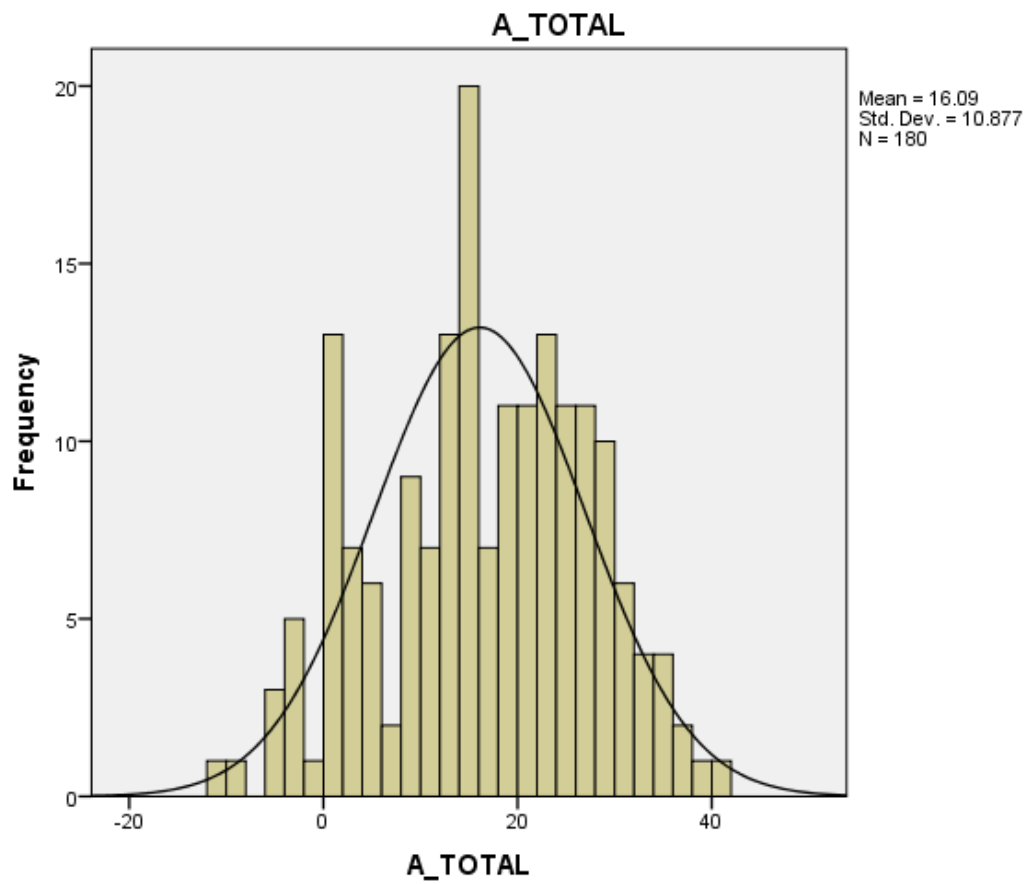
APPENDIX G

Histograms

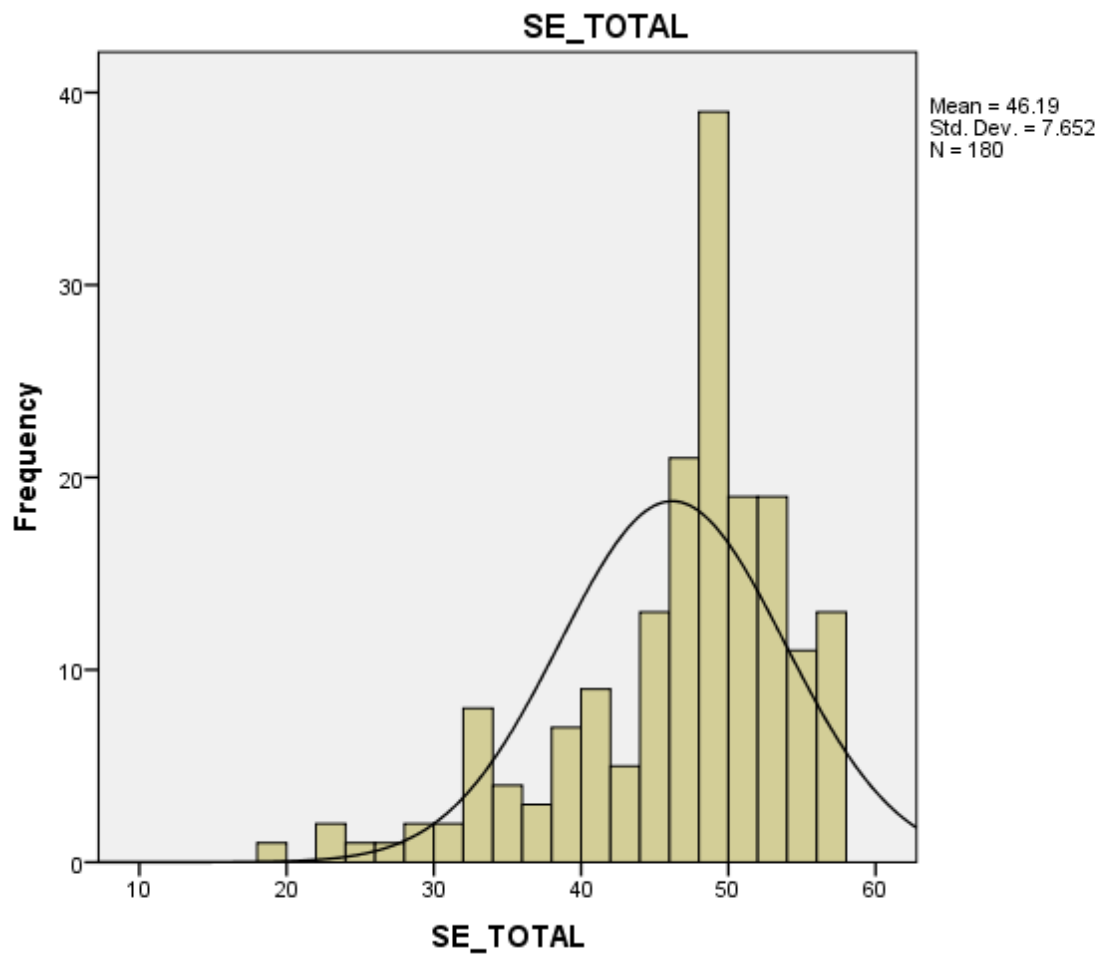
Conscientiousness



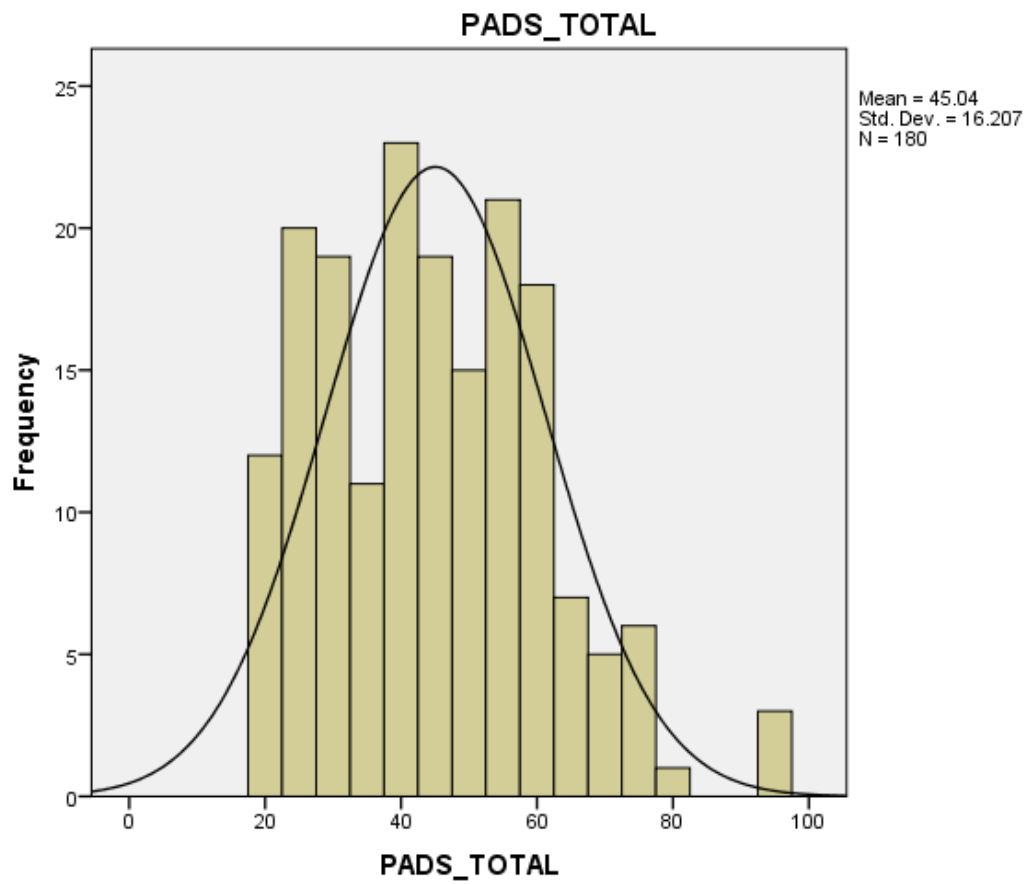
Agreeableness



Self-efficacy



Rater Discomfort



Performance Ratings

