FACTORS THAT AFFECT PEER RATER ACCURACY IN MULTIRATER FEEDBACK SYSTEMS

by

Katie J. Thiry

MICHAEL H. MCGIVERN, Ph.D., Faculty Mentor and Chair

KEITH J. JOHANSEN, Ph.D., Committee Member

JIM MIRABELLA, D.B.A., Committee Member

Barbara Butts Williams, Ph.D., Dean, School of Education

A Dissertation Presented in Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

Capella University

July 2009

Abstract

Multirater feedback, also referred to as multisource feedback or 360 degree feedback, has evolved into the performance appraisal system. The process of providing individuals with feedback from several sources, including coworkers, subordinates, clients, and supervisors, has emerged as a popular technique because it has the potential to provide a more complete picture of performance. Traditionally, the immediate supervisor was judged to be in the best position to evaluate an employee; however, using multiple rating sources leads to increased reliability, fairness and observational power as raters with differing perspectives and roles evaluate the employee (Harris & Scharbroeck, 1988; Latham 1989; Borman, 1997). Despite considerable research that has focused on the validity and accuracy of multirater feedback ratings compared to supervisory ratings, what remains ambiguous is a complete understanding of factors that may influence ratings provided by individuals who are not in roles traditionally known to evaluate employees, such as supervisors. Although multirater feedback holds promise as a means of evaluation that is less biased, more reliable, and more valid than the traditional supervisory appraisal method, no studies have focused specifically on the nonperformance factors that may influence the accuracy of peer ratings. Therefore, examining factors that have the potential to influence ratings provided by peers warrants attention. This study examined numerous nonperformance variables anticipated to affect peer responses to multirater feedback instruments. These findings may improve the interpretation and understanding of peer feedback in multirater feedback assessments.

Acknowledgments

I want to express sincere appreciation to my family, friends, colleagues, committee members, and fellow researchers, who supported and encouraged me through this process. Thank you, Jason D. Thiry, my husband; your commitment and dedication to my achievement of this goal was in many ways equal to my own. How blessed I am to have a partner who is as devoted to the dreams and ambitions of his spouse as he is to his own. Thank you, Daniel and Lisa Schneider, my parents; you have instilled in me the strength and fortitude to continuously set and achieve new goals. You lead by example.

I want to express sincere gratitude to the members of my dissertation committee. Your professional competencies are matched only by your infectious enthusiasm for teaching and mentoring. Thank you, Dr. Michael H. McGivern, my committee chairperson. First, and foremost, I am grateful that you agreed to undergo this journey and “adopt” me as your mentee. Thank you for your patience and support, from the start of my coursework, through the completion of this research. Thank you, Dr. Keith J. Johansen, for your willingness to share your knowledge and expertise during my coursework and your suggestions regarding this study throughout the research process. Thank you, Dr. Jim Mirabella, for sharing expert advice, generously investing your time, at all hours, and offering careful guidance and direction to ensure that this goal would be realized. You had confidence in me when my own wavered; for that, I am eternally grateful. I am fortunate to have developed lifelong friendships on this journey.

List of Tables

CHAPTER 1. INTRODUCTION

Introduction to the Problem

The use of performance appraisals is a widely used practice and serves a number of important functions within organizations. Each year in the United States over 70 million individuals receive some type of performance appraisal (Matens, 1999). Performance appraisals have traditionally been the responsibility of the supervisor or manager of the employee (Murphy & Cleveland, 1995). However, recent studies confirm that ratings from subordinates, customers, self, upper level managers, and peers have incorporated into the performance appraisal process, and supervisors are not the only source of ratings (Mohrman, Cohen, & Mohrman, 1995).

Multirater feedback (MRF), often referred to as multisource or 360-degree feedback, has emerged as a popular performance feedback method (Hedge, Borman & Birkeland, 2001; London & Smither, 1995). MRF is a system or process in which individuals receive confidential, anonymous performance feedback based on observations from different perspectives (Van Velsor, Leslie, & Fleenor, 1997). Throughout this document, the term multirater feedback (MRF) will be used to reference feedback based on observations from multiple raters. The person being rated will be referred to as the ratee and the person assigning a rating of performance will be referred to as the rater. Specifically, the term peer rater will refer to the person providing a rating of performance from the perspective of someone with similar expertise or level within the organization (i.e. co-worker). The term, 360-degree Feedback is a registered trademark of TEAMS, Inc. For the purposes of this research, the term multirater feedback will continue to be used to refer to the MRF process, itself, and not in reference to one specific product or assessment tool.

Not only is MRF aligned with today's business culture of work teams and less hierarchical organizations, it also involves the assumption, derived from measurement theory, that multiple raters yield more valuable information than any single individual (Church & Bracken, 1997).

The assumed advantages of MRF systems are their greater accuracy and objectivity compared to traditional top-down performance appraisals. For example, according to Bernardin (1992), peers are a valid source of performance information because peers work closely with the ratee; thus, they have many opportunities to observe the behavior of the person being rated. Consequently, peer ratings have the potential to provide more useful, valid data than supervisory ratings alone (Mohrman et al., 1995). The opportunity for peers to observe the ratee and provide feedback led Murphy and Cleveland (1995) to state, "All sources may have insights regarding an individual's strengths and weaknesses, but peers may represent the single best informed source" (p. 144).

Background of the Study

All performance appraisals consist of a performance rating system that requires raters to use their judgment, based on past observations, to measure and rate an individual's performance (Landy & Farr, 1980). Organizations use the results of these ratings for administrative purposes, such as personnel decisions, including salary increases, recommendation for promotion, job transfer, or developmental purposes, such as coaching or recommendations for training and development (Murphy & Cleveland, 1995).

Previous research indicates researchers' and practitioners' dissatisfaction with rater accuracy (Landy & Farr, 1980). Research has shown that only 20 percent of all appraisals considered are effective in assessing work performance (Matens, 1999). Performance appraisals are thought to be inherently biased because personal judgments, subjective values, and individual perception are fundamental to the process (Oberg, 1999). Biases and judgmental error introduce rating error in the evaluation of performance; thus directly affect the accuracy of the ratings.

As a result, a significant amount of existing research has examined the factors that contribute to the overall effectiveness of performance appraisals (Keeping & Levy, 2000). With the number of performance appraisals conducted annually and the role they play within organizations, it is logical that performance appraisal research would focus on the accuracy of ratings. The goal of previous research has been to establish "what factors other than actual performance of the ratee affect performance ratings and to determine methods by which these biases could be eliminated or minimized" (Wherry & Bartlett, 1982). It seems logical that research aimed at improving the "validity of judgmental measurements of performance" (Landy & Farr, 1980) is warranted and necessary. A great deal of research has focused on increasing the effectiveness of performance appraisals by improving rating format, designing techniques to improve long-term recall in raters, and by developing training programs to assist rater's recall of performance in ratees (Borman, 1997).

In a comprehensive literature review, two specific researchers, Wherry and Bartlett (1982), proposed that four broad factors influence the accuracy of performance ratings: the actual performance of the ratee, the observation of the performance by the rater, the rater's biases in the perception and recall of that performance, and measurement error.

Rater biases consist of two components – the rater's idiosyncratic tendencies and the raters' organizational perspective (London, 2001). The term 'idiosyncratic rater effects' is used to include all of the perspective-related effects associated with individual raters, such as the tendency to be lenient or harsh. The rater's organizational perspective (i.e. self, subordinate, peer, or boss) influences performance ratings for two reasons. First, raters from different organizational perspectives might observe different examples of the ratee's performance. For example, peers may have more opportunities to observe the performance of a co-worker on a day-to-day basis than a supervisor. Secondly, raters from different perspectives might observe the same aspects of performance but attach different attributes to them (Borman, 1997).

Previous research by Scullen, Mount, and Goff (2000) examined five components: general performance of the ratee, dimensional performance of the ratee, idiosyncratic rater effects, rater perspective effects, and random measurement error. Table 1 illustrates each of these effects for three rater perspectives within an MRF assessment: boss, peer, and subordinate. The table shows that the rater bias category is relatively large and accounts for the largest amount of variance (62%, averaged across perspectives). Within this category, the idiosyncratic component was the largest by far, accounting for over half (54%) of the total rating variance when averaged across boss, peer, and subordinate ratings, with the peer perspective accounting for the largest variance (58%). Organization perspective, the second component, was smaller than the idiosyncratic component for 8% of rating variance when averaged across perspectives.

Table 1. Percentage of Observed Rating Variance Associated With Five Categories of Effects for Boss, Peer, and Subordinate Ratings

	Perspective			
Categories of Effect	*Boss*	*Peer*	*Sub*	*Mean*
Rater bias effects				
Idiosyncratic effects (halo error)	47	58	57	54
Organization perspective	10	0	15	8
Total	57	58	72	62
Ratee performance effects				
General performance	19	21	10	17
Dimensional performance	11	8	5	8
Total	30	29	15	25
Measurement error	13	14	13	13

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*. 85, 956–970.

From the perspective of performance appraisal research, it is particularly noteworthy that the rater bias effects were the largest, by far. Within the rater bias category, idiosyncratic rater effects accounted for the largest amount of variance in ratings with 58% for peer ratings.

One result that differed across perspectives was the magnitude of the effect associated with organizational perspective. Meaningful effects were observed for subordinate (15%) and boss ratings (10%) but not for peer ratings (0%). According to London (2003), this means, “boss and subordinate raters attend to, encode, store, and retrieve social information about a ratee in ways that may be unique to raters from the same perspective. There is no evidence that this is true of peer raters” (p. 169).

Statement of the Problem

The basic assumption of performance ratings, including MRF ratings, is that they capture the performance of the person being rated. However, research results from Scullen, Mount, and Goff (2000) illustrate that about 25 percent of the variance in MRF ratings reflect ratee performance, whereas 54 percent represents the idiosyncratic tendencies of raters. According to Scullen et al., the MRF ratings indicate the rating tendencies of raters more than they measure the performance of the ratee.

A considerable amount of research has focused on the biases of raters. These biases produce unwanted variance in performance ratings. For example, during the rating process, a number of factors may influence rater judgment, and some of them may constitute potential sources of “error.” Among the potential sources of “error”, include halo and leniency, unintentional manipulation, and race, gender, or age biases (Facteau & Craig, 2001).

In order to improve rater accuracy, it is critical to identify and control bias. According to Wherry and Bartlett (1982), once biases are isolated and understood, methods can be developed to improve the assessment of performance. Despite considerable research that has focused on the validity and accuracy of MRF ratings compared to supervisory ratings, no studies have focused specifically on the nonperformance factors that affect the accuracy of peer ratings. This study addresses this gap by examining the influence of specific factors on rating accuracy in a peer rater feedback context.

Although MRF holds promise as a means of evaluation that is less biased, more reliable, and more valid than the traditional supervisory appraisal method, the research examined the extent to which MRF ratings, specifically peer ratings, were influenced by factors other than performance.

Purpose of the Study

The purpose of this study was to identify variables that may affect peer raters' responses to a multirater feedback instrument. Numerous variables anticipated to affect the peer rater's response to an MRF instrument have been examined. For example, this study adds to previous research by contributing to data on nonperformance characteristics (e.g. gender, age, etc.). Specifically, the relationship between raters' perception of the accuracy of their ratings to nonperformance variables such as: (a) gender, (b) age, (c) tenure, (d) concern for anonymity / confidentiality, (e) comfort with process, (f) opportunity to observe, (g) training, (h) purpose of assessment, (i) task association, (j) friendship, (k) likeability, (l) competition, (m) acquaintanceship, (n) rater selection process, and (o) favorability of the rating were included in this study.

This study has quantified the results of variables that may have affected the accuracy of an actual rating process as communicated by participants of this study. This study was designed to enhance understanding of peer performance ratings in the MRF process. For example, examining the variables that affect peer raters' responses to a multirater feedback instrument has provided insight into potential perspectives and bias that may influence the results of the performance evaluation. The findings of the study may improve the interpretation of peer feedback in MRF assessments. The intent is that this research will lead to additional research designed to improve MRF assessments, resulting in feedback that is more reliable, valid, and accurate. By investigating the factors that influence the accuracy of peer ratings, organizations can be more knowledgeable regarding proper utilization and the advantages and limitations of MRF for administrative and developmental purposes.

Rationale

Despite considerable existing research that has focused on the reliability, validity and accuracy of MRF ratings versus supervisory ratings, no empirical studies have focused specifically on the impact that potential perspectives and bias, such as interpersonal affect, have on the accuracy of peer ratings. Specifically, this study was designed to determine the possible factors that influence rater biases among peer raters in MRF. This study has explored rater issues that, to date, have not received adequate scrutiny. Given that many organizations utilize MRF methods, the results of this research should be of interest to practitioners. Because of the increasing use of peers as raters, there is a need for a better understanding of the reliability, validity, and bias of peer feedback.

Research Questions

The research questions used in this study have focused on the experiences and perceptions of peer raters who have participated in a multirater feedback system within their affiliated organization. To gain an understanding of the influence of factors that potentially affect the performance evaluations provided by peer raters, the following research questions were utilized for the focus of this study:

1. What is the relationship between the accuracy of multirater feedback ratings from peers versus the nonperformance factors / demographics of the rater and ratee?

2. What is the relationship between the accuracy of multirater feedback ratings from peers versus the personal relationship between the rater and ratee?

3. What is the relationship between the accuracy of multirater feedback ratings from peers versus the selection process for peer raters?

4. What is the relationship between the accuracy of multirater feedback ratings from peers versus the favorability of the overall rating?

The following hypotheses related to research question 1 have been tested:

Hypothesis 1a: Accuracy of multirater feedback ratings from peers is independent of the gender of the rater.

Hypothesis 1b: Accuracy of multirater feedback ratings from peers is independent of the gender of the ratee.

Hypothesis 1c: Accuracy of multirater feedback ratings from peers is independent of whether the rater and ratee are the same gender.

Hypothesis 1d: Accuracy of multirater feedback ratings from peers is independent of the age of the rater.

Hypothesis 1e: Accuracy of multirater feedback ratings from peers is independent of the difference in age between the rater and ratee.

Hypothesis 1f: Accuracy of multirater feedback ratings from peers is independent of the tenure of the rater with the organization.

Hypothesis 1g: Accuracy of multirater feedback ratings from peers is independent of the level of concern for anonymity and confidentiality of the rater.

Hypothesis 1h: Accuracy of multirater feedback ratings from peers is independent of the raters' comfort level with the multirater feedback process.

Hypothesis 1i: Accuracy of multirater feedback ratings from peers is independent of the opportunity raters had to observe the ratee.

Hypothesis 1j: Accuracy of multirater feedback ratings from peers is independent of the influence of rater training.

Hypothesis 1k: Accuracy of multirater feedback ratings from peers is independent of the purpose of peer ratings.

Hypothesis 1l: Accuracy of multirater feedback ratings from peers is independent of the familiarity of peer raters to the assigned responsibilities and tasks of the ratee.

Hypothesis 1m: Accuracy of multirater feedback ratings from peers is independent of nonperformance factors.

The following hypotheses related to research question 2 have been tested:

Hypothesis 2a: Accuracy of multirater feedback ratings from peers is independent of the friendship between rater and ratee.

Hypothesis 2b: Accuracy of multirater feedback ratings from peers is independent of how well liked the ratee is by the rater.

Hypothesis 2c: Accuracy of multirater feedback ratings from peers is independent of the degree of competition that exists between the rater and the ratee.

Hypothesis 2d: Accuracy of multirater feedback ratings from peers is independent of how long the rater has known the ratee.

The following hypothesis related to research question 3 has been tested:

Hypothesis 3: Accuracy of multirater feedback ratings from peers is independent of the rater selection process.

The following hypothesis related to research question 4 has been tested:

Hypothesis 4: Accuracy of multirater feedback ratings from peers is independent of the favorability of the overall rating.

Significance of the Study

Some researchers (e.g., London & Smither, 1995) have argued that research on MRF has not kept pace with practice and that there are insufficient research models and data available to guide organizations in the use of this type of feedback (Waldman & Atwater, 1998). By studying the factors that influence MRF, people in organizations can become more knowledgeable regarding the proper utilization, advantages, and limitations of this type of assessment. For example, examining how relationships affect raters' use of the rating scale across performance dimensions provides insight into perspectives and biases that may influence ratings. The findings of this study will contribute to the improvement of the interpretation of MRF.

Nature of the Study

This study included peer raters from a variety of industries and organizations. The individuals targeted for this study are members of the MN Chapter of the International Society of Performance Improvement (ISPI), members of the Front Range Chapter of the International Society of Performance Improvement (ISPI), members of the Seattle Chapter of the International Society of Performance Improvement (ISPI), members of the Puget Sound Chapter of the American Society of Training and Development (ASTD), and members of the Pacific Northwest Chapter of Organizational Development Network (ODN). These groups were selected because they offer a large sample size, include raters from a variety of industries and organizations, and a range of MRF assessment processes and tools. To be eligible for participation in the this study, participants must have completed the MRF assessment as a peer rater at least two weeks prior to their participation in the research study and within the most recent 18 months. Only those participants meeting the above criteria have been included as subjects for this study.

Assumptions and Limitations

Assumptions

Assumptions, which are those things the researcher holds as basic, obvious truths, can influence the selection of research methods and serve as the foundation of the overall research. Consequently, they can affect the validity of the study. One assumption made was that participants were willing to answer survey questions objectively and honestly. This study was based on the assumption that participants responded in a forthright way and did not purposefully attempt to be dishonest in one direction or another. It is likely that participants were truthful because the study was not implemented by or associated in any way with their affiliated organization. Additionally, participants, ratees, and their organizations remained anonymous and individually unidentifiable.

Another assumption was that participants would remember the rating that they provided in a previous peer assessment. It was likely that participants would recall the experience because the act of participating in a peer assessment process occurs infrequently (i.e. annually) and the participants were prompted to recall assessment feedback from and within the preceding 18 months. It was also assumed that the number of participants used in this study was adequate and appropriate.

Limitations

The limitations of a study are those areas and elements that cannot be controlled sufficiently; nor can they be explained when conducting the analysis, interpretation, and generalization of the data collected. One limitation of this study was that individuals often have strong emotions associated with the term “performance appraisal” or "performance evaluation." Therefore, their responses may have included a degree of emotion that is immeasurable and a potential threat to validity. Additionally, questionnaires, as were used in this study, are limited in their ability to probe deeply into responses, beliefs, attitudes, and inner experiences (Gall, Gall, & Borg, 2003). Once the questionnaire was distributed to the participants, the researcher was not able to modify the items, even if they were unclear to some respondents (Gall et al., 2003). To control this limitation, the questionnaire was designed to include a comments field to allow participants to make clarifications of their responses and comment, as appropriate. Despite these limitations, the research results will be valuable to a wide variety of industries and organizations using MRF.

Definition of Terms

This document references a number of terms. The terms used in this document are defined below:

1. *Administrative purposes*. The information used to make decisions about pay, promotion, access to resources, termination, etc., is defined as administrative (Wiese & Buckley, 1998).

2. *Bias*. The systematic tendency for ratings to be influenced by anything other than the behavior being measured is referred to as bias (Landy & Farr, 1980; Bigoness, 1976).

3. *Developmental purposes*. The focus of the information used for the future improvement or personal and/or professional development to improve current job performance, knowledge or skill levels of an individual defined as developmental (Burke, Weitzxel, & Weir, 1978; Dorfman, Stephen, & Loveland, 1986; Meyer, Kay, & French, 1965).

4. *Feedback*. The information sent or received about an individual's performance is referred to as feedback (Hillman, Schwandt, & Bartz, 1990).

5. *Interpersonal Effect/Rater Affect*. The tendency of an assessor or an evaluator to rate an individual's performance at a level that does not accurately or consistently reflect the performance level of that individual is referred to as interpersonal effect or rater effect (Antonioni, 1994). There are several types of rater affect, all of which are possible sources of error in measurement and rating, which will be reviewed within this document.

6. *Multirater, multisource, multilevel, and 360-degree feedback systems*. Each of these terms refers to formal or informal procedures for obtaining performance information from more than one individual and/or more than one level within the organization (Harris & Scharbroeck, 1988; Latham, 1989; Borman, 1997).

7. *Peer*. Someone with similar expertise or level within the organization (i.e. co-worker) will be referred to as a peer.

8. *Peer feedback*. The information provided by someone with similar expertise or level within the organizational hierarchy as the individual being evaluated or assessed is referred to as peer feedback (Kane & Lawler, 1978).

9. *Peer rater*. The person providing a rating of performance from the perspective of someone with similar expertise or level within the organization as the individual being evaluated or assessed (i.e. co-worker) is referred to as a peer rater (Wherry & Bartlett, 1982).

10. *Performance appraisal, performance evaluation, performance review*. These terms refer to the data and/or information collected as part of a formal evaluation process and used for developmental and/or administrative purposes.

11. *Ratee*. The individual being evaluated or assessed will be referred to as the ratee.

12. *Reliability*. Reliability is "the extent to which a set of measurements is free from random-error-variance" (Guion, 1965, p. 30).

13. *Validity*. Validity is the extent to which an instrument measures what it is supposed to measure (Guion, 1980).

Organization of the Remainder of the Study

The following is an outline of the structure of this research study. Chapter 1 introduced the problem, the purpose and rationale of the study as well as the specific research questions and hypotheses that were tested. These sections were followed by the assumptions and limitations specific to this study.

Chapter 2 presents a summary of the literature related to peer evaluations in multirater feedback assessments, the history of performance appraisal, trends in performance appraisal, benefits of multirater feedback, and issues surrounding multirater feedback. The final section summarizes the significant issues discovered in the review.

Chapter 3 begins with an overview of the research methods, participants, and data collected to answer the research questions. This chapter provides a detailed description of the participants in the study, the research methods and instruments used, the procedures to be used to collect the data, how confidentiality was maintained, and the questions included in the research questionnaire.

Chapter 4 reports the data analysis and results of the study. The results of the hypotheses testing are introduced and followed by an analysis of the results.

Chapter 5 explores the results, conclusions and recommendations resulting from the study. The research questions and supporting hypotheses followed by a summary of conclusions based on the results of the study and presented. Recommendations for future research complete the presentation of this chapter.

CHAPTER 2. LITERATURE REVIEW

Introduction

The chapter begins with an overview of the history of performance appraisal and trends, following an introduction to the need for individual feedback, and a review of the benefits of multirater feedback. The chapter ends with a review of the literature connecting the issues of peer feedback to the larger context of performance appraisal.

History of Performance Appraisal

Although the practice of formal evaluation has existed for centuries, today's performance appraisal began with the research by industrial/organizational psychologists at Carnegie-Mellon University on the use of "man to man" forms to select salespersons (Scott, Clothier, & Spriegal, 1941). Similar forms were used by the Army during World War I to assess the performance of officers (Scott et al., 1941). After World War I, many of the psychologists employed by the military began working in industry; as a result, the popularity of performance appraisals increased dramatically. By the early 1950s, performance appraisal was an accepted and common practice in organizations. In fact, a 1962 survey found that performance appraisals were already conducted in 61% of organizations (Spriegal, 1962). Although during this time, performance appraisal was usually limited to employees at the bottom and middle of the organizational hierarchy (Whisler & Harper, 1962).

Trends in Performance Appraisal

In recent years, multirater feedback has emerged as a popular performance feedback method (Hedge, Borman & Birkeland, 2001; London & Smither, 1995;

Romano, 1994). Traditionally, performance appraisals have been the responsibility of the supervisor or manager of the employee (Murphy, Cleveland, & Mohler, 2001). Recent studies confirm that supervisors are not the only source of ratings in organizations; and additional sources include subordinates, customers, self, upper level managers, and peers (Mohrman et al., 1995). Not only is multirater feedback aligned with today's business culture of work teams and less hierarchical organizations, it also involves the assumption, derived from measurement theory, that multiple sources yield more valuable information than any single individual (Church & Bracken, 1997). Although the idea of using peers in the performance appraisal process is not new, practitioners have recently demonstrated an increased interest in peer evaluation (Kanter, 1989; Peterson & Hillkirk, 1991).

Antonioni (1994) reported that 25% of companies use some form of MRF assessment process. Another report indicated that as many as 12% to 29% of all U.S. organizations were using this method (Bracken & Church, 1997). Timmreck and Bracken (1995) cite a survey of a corporate consortium whose 20 large companies routinely utilize MRF assessments. The survey results indicated that over half of them use MRF assessment company-wide. The data collected from the MRF assessments is used for development and coaching in 93% of the companies, with 28% utilizing it as input for appraisal. A majority (56%) of these organizations conduct MRF assessments annually. According to research detailed by Lepsinger and Lucia (1997), every Fortune 500 firm is either doing it or thinking about doing it. An estimated 25% of all organizations use some type of 360-degree feedback as a tool for leadership development and 90% of all organizations use MRF as a part of his or her performance management system (Nowack, 1993). Clearly, the use of MRF has not diminished, and perhaps has increased. With the increase in team-based structures in organizations, feedback instruments designed specifically for team members to receive feedback from one another about their team behaviors and performance will become increasingly popular.

Overview of Multirater Feedback

It is important to note that the terms multirater feedback, 360-degree feedback, and multisource feedback are used interchangeably to describe assessments involving ratings from multiple evaluators (Hedge, Borman, & Birkeland, 2001). Evaluations of individual performance through self-ratings, peer ratings, supervisor ratings, and subordinate ratings are based on the philosophy that individuals should receive a full (i.e., 360 degree) picture of their performance from multiple perspectives. Typically, raters are grouped by organization level, including subordinates, boss, peers, and the individual being assessed. Several research studies show that the value of multirater feedback is its ability to provide a rich source of performance feedback from individuals who have unique viewpoints. According to Tsui & Barry (1986), multiple sources are necessary because a lack of agreement often occurs when assessing overall performance.

The MRF process usually begins with a combination of about eight to twelve people who complete an anonymous survey that asks questions covering a broad range of workplace competencies. The surveys generally include questions that are measured on a rating scale and ask raters to provide written comments. The person receiving feedback also fills out a self-assessment that includes the same survey questions that others receive. Individual responses are typically combined with responses from other people in the same rater category (e.g. peer, direct report) in order to preserve anonymity and to give the ratee a clear picture of his or her greatest overall strengths and weaknesses.

Multirater Feedback Uses

Multirater feedback systems are most typically used as a development tool to help employees recognize strengths and weaknesses and as a performance appraisal tool to measure employee performance. When done properly, MRF is highly effective as a development tool. For instance, the feedback process gives people an opportunity to provide anonymous feedback to a coworker that they might otherwise be uncomfortable giving. Feedback recipients gain insight into how others perceive them and have an opportunity to adjust behaviors and develop skills that will enable them to excel at their jobs. MRF is also used as a component for selecting courses for development based on the identified needs of the individual. The most popular application for MRF is for management development. Management development experiences focus on enhancing a person’s leadership capabilities. When MRF is used for management development purposes, the common practice is that no one in the company sees the feedback report with the exception of the participant and a neutral party who processes and distributes the results.

Using a MRF system for performance appraisal is also a common practice. Some companies link MRF results to administrative decisions, such as performance appraisal, compensation, succession planning, or promotions.

Rater Research

Previous studies have been conducted on rater/ratee characteristics to determine if the interactions between rater and ratee result in measureable biases. This broad area of research has focused on analyzing performance ratings as a function of rater and ratee demographic characteristics. This research has attempted to isolate inaccuracies or variation caused by nonperformance factors. The majority of this research has centered on how gender, race, or age bias affected rating accuracy (Hartel, Douthitt, Hartel, & Douthitt, 1999; Landy & Farr, 1980; Bigoness, 1976; Hamner, Kim, Baird & Bigoness, 1974; Pulakos & Wexley, 1983; Nevill, Stephenson, & Philbrick, 1983).

Landy and Farr (1980) believe that these researchers have too narrowly focused their studies by looking at too few demographic characteristics or by just concentrating on either the rater or the ratee characteristics alone. They believe that unmeasured or hidden variables may have an unknown effect on the results of these previous studies (Landy & Farr, 1980). While the interaction between rater and ratee demographic characteristics has been studied, it has been limited to the effects of race or gender (Mobley, 1982; Pulakos & Wexley, 1983; Schmitt & Lappin, 1980; Landy & Farr, 1980). However, the results of these studies seem to indicate that there is a positive relationship when rater and ratee race are similar. The results of rater and ratee gender research are more mixed but there also seems to be indications that male raters rate female performance lower than they do other males (Landy & Farr, 1980).

Research findings by Zalesny (1986) indicate that rater/ratee differences in specific characteristics might have an even greater effect on ratings than the effect of being similar. Others found that when raters perceive similarities between themselves and the ratees that perception has an even greater effect on performance ratings than the existence of actual similarities (Strauss, Barrick, & Connerly, 2001; Turban & Jones, 1988). These findings indicate a possible source of bias.

This research study has added to the findings of these previous studies. The purpose of this research was to identify potential sources of bias within peer ratings in hopes that, once identified, the biases can be eliminated or reduced. The discovery of these sources of bias will ultimately improve the effectiveness of the peer feedback.

Benefits of Multirater Feedback

Researchers have suggested that the advantages of using multiple raters include the ability to observe and rate various job facets of each ratee’s performance (Borman, 1997), greater reliability, enhanced fairness, and increased ratee acceptance (Wexley & Klimoski, 1984). Previous empirical research has addressed the benefits of multirater feedback (London & Beatty, 1993; Tornow, 1993), the benefits of peer and upward appraisals (Fedor, Bettenhausen, & Davis, 1999), and the extent of self-other agreement in ratings (Atwater, Ostroff, Yammarino & Fleenor, 1998; Atwater, Roush, & Fischthal, 1995; Atwater & Waldman, 1998). Considerable evidence suggests that peers can be reliable and valid predictors of job performance (Kane & Lawler, 1978; Lewin & Zwany, 1976; Schmitt, Gooding, Noe, & Kirsh, 1984). Namely, peer evaluators often work closely with the people whom they assess; naturally, they have access to a wider range of performance dimensions (Borman, 1997) and may be able to make more precise performance distinctions across ratees. Not only are peer appraisals likely to be based on different, perhaps more accurate information than supervisory appraisals, but the group influence literature predicts that they might be more effective than supervisory appraisals in producing behavioral changes (Festinger, 1954).

Other advantages of multirater feedback systems are their greater accuracy and objectivity compared to traditional top-down performance appraisals. According to Bernardin (1992), peers are a valid source of performance information. Peer ratings have the potential to provide more useful, valid data than supervisory ratings alone (Mohrman et al., 1995). The opportunity for peers to observe the ratee and provide feedback led Murphy and Cleveland (1995) to state, “all sources may have insights regarding an individual’s strengths and weaknesses but peers may represent the single best informed source” (p. 144).

Shaver (1995) suggests that MRF helps people “uncover expectations, strengths, and weaknesses that are news to them; it broadens the perspective on evaluating an individual by using multiple data sources; it provides ratings that can become benchmarks in the feedback recipient’s performance appraisal process; it may promote people becoming increasingly accountable for their own growth and development; and it is an efficient procedure in that it is inexpensive, simple, and quick” (p. 13).

The Role of Feedback

The act of providing feedback is “the activity of providing information to staff members about their performance on job expectations” (Hillman, Schwandt, & Bartz, 1990, p. 20). Feedback plays an important role in that it is not only the information people receive about their performance, but feedback “conveys an evaluation about the quality of their performance…" (London, 2003, p. 11). Previous research on feedback indicates a number of reasons why feedback is so important to enhancing work performance. Based on literature reviews by Ilgen, Fisher, and Tayloar (1979), Larson (1984), London (1988), and Nadler (1979), feedback directs behavior, influences future goals, reinforces positive behavior, and heightens an individual's sense of achievement and internal motivation.

Meaningful feedback is central to performance management. As London (2003) observed, "Psychologists have long recognized the value of feedback to enhance job challenge, increase motivation, and facilitate learning when the information is meaningful and given in a helpful way" (p. 3). Feedback guides, motivates, and reinforces effective behaviors and reduces or stops ineffective behaviors. However, givers of feedback may be intentionally or unintentionally biased, destructive, or hurtful. Specifically, raters may be swayed by factors unrelated to actual performance and, as a result, may offer useless information. According to Waldman and Atwater (1998), "Feedback has been found to increase the accuracy of self-perception, as well as to give individuals information about how others perceive their behavior" (p. 5).

Issues Surrounding Multirater Feedback

Is there no reason to believe that multirater systems prevent many of the rating errors and distortions found in traditional appraisals? For example, peer feedback may be compromised by the practice that ratees often select their raters. Ratees may be inclined to select friends as raters, and friendship may lead to positively enhanced assessments. In other words, is the method of using peer raters introducing new biases into the evaluation process? On the other hand, some researchers have suggested that peer raters may be negatively biased because they are competing for the same organizational rewards as ratees (DeNisi & Kluger, 1996). Even if they are not competing for the same organizational rewards, peer appraisers could be viewed as negatively biased for other reasons. These, and other nonperformance factors, have been explored in this research study.

Rating Biases and Common Rating Errors

In a traditional employee performance appraisal, the sole rater was the immediate supervisor. However, the primary reason traditional performance appraisals have difficulty in yielding accurate results is because the supervisor often lacks sufficient information concerning the individual’s actual performance (Longnecker & Goff, 1990). Additionally, supervisors may not have sufficient information to accurately evaluate an employee’s behavior. The resulting rating is based on impressions, which may lead to errors and biases (Longnecker & Goff, 1990).

A considerable amount of research has focused on the biases of raters, in general. During the rating process for self and others, a number of factors influence a rater’s judgments, and some of them may constitute potential sources of “error.” Rating errors reduce reliability and validity when inaccurate results are gathered (Roberts, 1998).

Despite their advantages, peer evaluations can be scrutinized because of concerns that the judgments of others’ performance may be affected by a variety of perceptual errors. Research in psychology helps to explain how people process information about others. The accuracy of interpersonal perceptions is important to consider when referring to feedback from others. Rater motivation, observation skills, information distorting biases, and empathy for others all influence rater accuracy (London, 2003). The social psychological processes of person perception explain how individuals form impressions of others and use this to provide them with feedback about job performance (London, 2003, p. 52). This concept refers to the process by which we form impressions and make inferences about other people in response to behaviors, words, and interactions observed (Klimoski & London, 1974). An evaluation of individual performance is subject to numerous factors that affect the accuracy and usefulness of the opinion made. Klimoski and London (1974) suggest that person perception incorporates the perceiver, the individuals perceived, the relationships between them, and the situation. The goals, motivation, and cognitive skills and processes of the perceiver must also be considered.

These common biases often are evident in performance ratings: (a) halo error/effect, (b) similarity error, (c) central tendency, (d) leniency, (e) harshness, (f) first impression, (g) reliance on stereotypes, and (h) fear of retaliation.

1. *Halo error/effect.* The halo error can be described as the tendency to allow perceptions of one performance dimension to influence ratings of other, unrelated performance dimensions. This is the tendency to rate a person the same or almost the same on all items. For example, if a rater thinks that the individual is highly competent in one area, the rater may rate that individual high on many other competencies as well. Whereas, the reverse is called halo effect. This describes the tendency of a rater to think the individual is not competent in one area, so the rater may rate that individual low on many other items (Landy & Farr, 1980).

2. *Similarity error.* Similarity is the tendency to give overly favorable ratings to ratees who are similar to the rater in characteristics unrelated to performance (e.g., age, race, or gender) (Fiske, 1993).

3. *Central tendency.* Central tendency is the tendency to give midrange ratings of all performance dimensions regardless of actual performance (e.g., ratings of 3 on 1 to 5 scales) (Landy & Farr, 1980).

4. *Leniency.* Leniency is the tendency to give mostly high or overly favorable ratings on all performance dimensions regardless of actual performance (Bracken, Timmreck, & Church, 2001). Leniency, or friendship bias, is a particular concern among peers completing peer ratings (Love, 1981).

5. *Harshness.* Harshness, on the other hand, is the tendency for some raters to be severe in their judgments. This is the tendency to give mostly low or overly negative ratings on all performance dimensions regardless of actual performance (Bracken, Timmreck, & Church, 2001).

6. *First impression.* First impression is the tendency to allow one's first impression of the rate to influence ratings.

7. *Reliance on stereotypes.* A reliance on stereotypes develops among peer raters when a new cohort is added to the peer group. This allows peers to make ratings after only having a limited amount of time to observe the new person's performance. Interestingly, this reliance on stereotypes is probably one reason why peer ratings are so stable over time (Passini & Norman, 1966). That is, stereotypes provide a common frame of reference for peers when rating.

8. *Fear of retaliation.* Finally, the fear of retaliation can be a real problem among peer raters. Raters who have received low peer ratings have retaliated against those peers during later rating opportunities (DeNisi & Kluger, 1996). For example, Koeck and Guthrie (1975) found that people gave lowered ratings to peers that they believed had given them low ratings during an earlier rating process. Additionally, the credibility of multirater feedback results may be in question if raters have a stake in the results, as they would if their co-worker's annual bonus depends in part on the ratings of peers who want their co-worker to be treated well in hopes that they, in turn, will be treated well. Additionally, individuals may try to influence how their peers rate them by implying a request for positive results.

However, there is a misconception that the raters must agree on their perceptions of an individual in order for their data to be reliable. There may be honest, stable, real differences in perceptions, based on different observational sets of the raters. Raters themselves have trouble with the concept of peer ratings for these reasons. In addition, if peer raters are chosen by the ratee, they may not select impartial people. Some contend that peer appraisers will be perceived as negatively biased because they are competing for the same organizational rewards as the ratees (DeNisi & Kluger, 1996). Unfortunately, feelings and judgments can complicate the act of providing useful feedback. After all, interpersonal feedback is inherently subjective.

No guarantees exist that, when provided the opportunity, raters in a multirater feedback process will provide good feedback. Feedback that lacks quality cannot benefit the recipients and, thus, will less likely benefit the greater organizational culture.

> Bad feedback has several characteristics. First the actual ratings are fraught with rating errors, such as central tendency, or using only the middle values of the rating scale; leniency, which can be both positive and negative; and halo. Second, the ratings contain biases, such as game playing that can occur when 360-degree feedback is highly evaluative within a culture lacking trust. In both these cases, this inaccurate feedback can be worse than no feedback. Third, the feedback provided in a 360-degree process can be qualitative in nature, especially if surveys involve writing-in comments. Such comments do not help if provided in very general terms. Bad feedback can also stem from bad survey items that are too general or given to raters unfamiliar with a ratee's behavior in the areas being rated (Waldman & Atwater, 1998, pp. 111-112).

Legal Implications Surrounding Feedback

As London (2003) suggests, "The performance review process must be conducted in a professional and fair manner, focused on behaviors and outcomes (not personalities) and free of discrimination unrelated to job performance" (p. 5). Title VII of the 1964 Civil Rights Act, the Civil Rights Act of 1991, and the Age Discrimination in Employment Act of 1975, works to protects individuals from age, race, religion, gender, or national origin discrimination. Because appraisals are subject to raters' subjective biases and prejudices, the legal implications of an appraisal system that depends on subjective criteria and personality traits should be a subject of concern.

Peer Evaluation

Peer ratings refer to rating of performance from an individual’s co-workers. Managers are not always available to observe all aspects of their subordinates’ performance, and peers have been regarded as being more knowledgeable about co-worker performance because of their day-to-day interactions (Druskat & Wolff, 1999; Fedor, Bettenhausen, & Davis, 1999). Increased reliability comes from peers raters’ who can view performance on a regular basis (Murphy & Cleveland, 1995).

Methods of Peer Evaluation

There are three basic methods of peer assessment: peer nomination, peer rating, and peer ranking (Kane & Lawler, 1978). Peer nominations consists of having each member of a group select a member or members as possessing the highest standing on a rating dimension. Conversely, the group members are asked to select the member or members that have the lowest standing on the same rating dimension. Peer rating consists of having each member rate each member of the group on a given set of performance characteristics. In some peer rating examples, evaluators are directed to allocate a specific total of points among group members referred to as “forced peer ratings.” Peer ranking is the method of each member of the group ranking all of the other members of the group from high to low on a set of performance characteristics.

Advantages of Peer Evaluation

Research suggests that understanding peer ratings is slightly less straightforward than other rating sources. The motivations of peers or team members can range from competitive to supportive to brutally honest, depending on the climate of the group and how the feedback is to be used. In spite of these complications, research shows that peers observe more examples of work behavior across a variety of situations and that their ratings are a better predictor of who will be promoted than any other rating source (Edwards & Ewen, 1996). Peers are likely to be effective raters of communication skills, interpersonal skills, decision-making ability, technical skills, and motivation (Brutus, Fleenor, & London, 1998).

The inclusion of peer ratings can be a positive and a negative. A study by DeNisi, Randolph and Blencoe (1983) demonstrated that knowledge of negative peer-ratings feedback resulted in lower performance, cohesiveness, and satisfaction on a task. However, a trend toward higher values on these variables was found for knowledge of positive peer-rating feedback. In a study done by Dominick, Reilly and McGourty (1997), positive peer ratings were shown to lead to an increase in group performance on a task. People are also more likely to be accepting of peer ratings when they are used for developmental rather than administrative purposes (McEvoy & Buller, 1987). Evidence also suggests that peers are more comfortable in their role as raters when the evaluation is being used for developmental rather than administrative purposes (Murphy & Cleveland, 1995).

Overall, however, peer ratings have been shown to be reliable and valid measures of managerial performance (DeNisi & Kruger, 1996; Love, 1981; Reilly & Chao, 1982). Murphy and Cleveland (1995) suggest three main reasons for this. First, peers work closely with the rate and have more opportunity to view their task performance as well as their interpersonal behaviors. This is consistent with Wherry and Bartlett’s (1982) theory of performance rating, which states that rater-ratee proximity is a key component to rating validity. Second, the presence of peers is less likely to induce censoring of behaviors than the presence of a supervisor. Therefore, it is more likely that peers see a less biased view of the ratees' performance. Third, the use of peers allows the pooling of ratings, which helps increase reliability. In this manner, the impact of any inconsistent ratings is reduced. This is not to say that peer ratings are free from criticism.

Rating Scales

Rating scales are used to capture raters' perceptions about whether, or how well, the individual being rated demonstrates the surveyed behaviors and skills. Most scales associate numbers with anchors; for example, 1 to 5, where 1 = Strongly Disagree, 5 = Strongly Agree. These are used to compute a numerical score. Some scales use only verbal descriptors, such as Strongly Agree and do not associate the verbal rating with a numerical value; these descriptors are later converted into numerical values for reporting purposes.

Scales can differ in the number of points and the number of choices that are included. Generally, scales range from three to 15 points. Most multirater feedback designers use a five-point scale, or they use four to six points so that there is no middle point. By eliminating a middle point, survey designers overcome the problem of the raters' propensity to overuse the safest choice on the scale, the middle or average rating.

It is often debated whether to include a Not Applicable (NA) or Don't Know (DK) rating choice. The rationale here is that raters need to be able to distinguish items that are not relevant or that they have not observed. The advantage to including NA or

DK as a rating choice is that these choices are not computed in the item's average score. When there is no NA or DK, raters often choose the middle point of the scale to express Not Applicable or Don't Know; this can lead to confusion about what the middle point actually represents.

Conclusion

Traditionally, performance appraisals have been the responsibility of the supervisor or manager of the employee (Murphy & Cleveland, 1995). Today, multirater feedback is a popular performance feedback method (Hedge, Borman, & Birkeland, 2001; London & Smither, 1995; Romano, 1994). While researchers have suggested that the advantages of using multiple raters include the ability to observe and rate various job facets of each ratee's performance (Borman, 1997), greater reliability, enhanced fairness, and increased ratee acceptance (Latham, 1989), the findings of previous research indicate possible sources of bias. A considerable amount of this research has focused on the biases of raters, in general. For example, during the rating process, a number of factors influence a rater's judgments, and some of them may constitute potential sources of "error." Rating errors reduce reliability and validity when inaccurate results are gathered (Roberts, 1998). Thus, examining the influence of nonperformance factors that may affect peer rater responses in a multirater feedback system warrants attention.

CHAPTER 3. METHODOLOGY

Introduction

This chapter describes the methodology used to identify, understand, and quantify potential factors that affect the accuracy of performance ratings given by peers. It will address: (a) the purpose of the study, (b) rationale for using the survey method as the research design, (c) research questions, (d) population and sampling procedures, (e) instrumentation, (f) reliability and validity, (g) ethical considerations, (h) the process used for data collection, and (i) data analysis procedures.

Conscious and unconscious biases can induce unwanted variation in performance ratings. There is a lack of empirical research on understanding the source and magnitude of these biases. Both quantitative and qualitative research methods were used to accomplish this analysis. A qualitative analysis was used to augment information obtained in the literature review about potential bias. This study quantified the nonperformance factors reported by peer raters that affected the accuracy of performance ratings made in a MRF system. Data for the study came from real individuals from real organizations obtained after having participated in an actual MRF process within their affiliated organization.

Previous studies on the effect of nonperformance factors on rating appraisals have generally examined only one or a few variables at a time. The results of these studies were questioned because of the probability that additional factors that were not controlled by the study had an unmeasured effect on the study’s results. This study avoided this criticism by including as many measurable nonperformance variables as possible. Each of these variables are discussed in this chapter.

Research Design

A mixed methods research design was used to investigate the research questions and hypotheses. A quantitative analysis was used to analyze the research questions from a numerical point of view to examine the variables among a sample population in order to make assumptions about peer raters participating in a MRF process as a whole. This methodology helped to minimize researcher bias because numerical data guided the analysis.

A qualitative analysis augments information in the literature review about potential bias. Specifically, data from focus groups was used to evaluate the variables proposed by the researcher and provide preliminary data for appropriate modification to the survey questionnaire used in this study. Results provided generalizations about assessments provided by individuals who participated in multirater feedback systems as peer raters.

Survey research is used to describe the research method that involves administering questionnaires (Gall, Gall, & Borg, 2003). The research method involves administering an electronic survey “to collect data about phenomena that are not directly observable: inner experience, opinions, values, interests, and the like” (Gall et al., 2003, p. 222). “The purpose of a survey is to use questionnaires or interviews to collect data from a sample that has been selected to represent a population to which the findings of the data analysis can be generalized” (p. 223). Survey research promotes understanding of a phenomenon, and given the goals of this study, it is an appropriate research method. This survey method allows for researchers to quickly and efficiently gather information from a larger group of study participants (Creswell, 2003; Gay, Mills, & Airasian, 2006). Gall et al. (2003) suggest that survey research involves these steps: (a) defining research objectives, (b) selecting a sample, (c) designing the questionnaire, (d) pilot-testing the questionnaire, (d) precontacting the sample, (e) writing a cover letter, (f) following up with nonrespondents, and (g) analyzing the questionnaire data. “The purpose of a survey is to use questionnaires or interviews to collect data from a sample that has been selected to represent a population to which the findings of the data analysis can be generalized” (Gall et al., 2003, p. 223). An electronic survey questionnaire was used in this study to determine how participants responded to a MRF assessment as peer raters. The survey questions were designed to address the research questions and provide the necessary data.

Research Questions

To gain an understanding of the variables anticipated to affect peer responses to a multirater feedback instrument, the following research questions were be examined:

1. What is the relationship between the accuracy of multirater feedback ratings from peers versus the nonperformance factors / demographics of the rater and ratee?

2. What is the relationship between the accuracy of multirater feedback ratings from peers versus the personal relationship between the rater and ratee?

3. What is the relationship between the accuracy of multirater feedback ratings from peers versus the selection process for peer raters?

4. What is the relationship between the accuracy of multirater feedback ratings from peers versus the favorability of the overall rating?

These questions provided the foundation upon which the research methods were selected. Survey questions were developed to correspond with each of these research questions (Appendix).

Population and Sampling Procedures

Focus Groups

Focus groups are carefully planned discussions designed to obtain perceptions on a defined area of interest (Krueger, 1988). Focus groups are group interviews in which the researcher relies on in-group interaction and discussion, based on topics that are supplied by the researcher who takes the role of a moderator (Morgan, 1997). The results from focus groups served as a source of data for the development of the survey questionnaire in the quantitative aspect of this research (Frey & Fontana, 1991). The decision to use the focus group method in this study was driven by the desire to gather a breadth of information from the research participants. According to Blumer (1969), during the exploratory phase of data collection, "a small number of individuals, brought together as a discussion or resource group, is more valuable many times over than any representative sample" (p. 41).

According to Morgan (1997), the hallmark of focus groups is their explicit use of group interaction to produce data and insights that would be less accessible without the interaction found in a group. Krueger (1988) agrees, and supports that focus groups work because attitudes and perceptions develop, in part, by interaction with others. As he states, "We are a product of our environment and are influenced by people around us" (p. 23). Lofland and Lofland (1984) noted that an advantage of focus groups are that they allow people more time to reflect and recall experiences and “something that one person mentions can spur memories and opinions in others” (p. 15). Therefore, focus groups served as an efficient and appropriate research technique in this research study.

Regarding the “ideal number” of focus groups to conduct, most researchers agree that three to five groups are usually adequate, as more groups seldom provide new insights (Morgan 1997; Krueger 1988). However, the final number of focus groups conducted should actually reflect the research plan (Bloor, Frankland, Thomas, & Robson, 2001). In their discussion of some of the “unusual problems” with group interviews, Fontana and Frey (1994) note that there is the possibility of one person or a small group of persons dominating the discussion while others will not speak up. This issue is largely associated with the size of the focus group. Research indicates that group size is central to the success of the focus group method. However, opinions vary regarding the “ideal size” for a focus group, with the literature pointing to an optimal number of 8-10 participants (Frey & Fontana, 1991) or 6-12 participants (Morgan, 1997). While groups have been reported as small as three participants to groups as large as 20 (Morgan, 1997). Based on recommendations of prior research, this research plan included two focus groups of 6-12 participants selected from senior management level employees from a large privately owned manufacturing facility.

Recruitment

For this study, the approach for recruiting focus group participants was researcher-driven. This research was not supported by a specific organization or company, therefore, the researcher was solely responsible for recruiting the research participants. These focus group participants were selected from a pool of exempt-level employees, with managerial responsibilities, who have all participated in an annual multirater assessment as peer raters. All of these raters have completed an online survey, using a 1 to 5 scale to rate the focal manager on a series of items related to leadership or management effectiveness. The focal manager received a report of results from a corporate trainer.

At the request of the researcher, the participants were identified by an executive who personally knew the individuals and assessed them to have adequate experience with MRF to provide useful input. Participation in the focus group was voluntary and participants were solicited through electronic mail. Each of the participants had completed a MRF assessment as a peer rater within the previous 18 months.

Procedures

Two focus groups were led by the researcher. Each focus group session began with the researcher welcoming the participants and thanking them for their time. Next, the researcher briefly described the purpose of the focus group and the research project, and then explained that the interview would last approximately 45-60 minutes, that it would be recorded for transcription purposes only, and that all names would be kept confidential. A brief description of the facilitator’s role, the participants’ roles, and the ground rules for participation was reviewed. The interviewees were reminded that all members of the group should be allowed to participate equally and that only one person should speak at a time. Open-ended questions were used in the interview and each focus group interview began with a “warm-up” question to initiate and encourage discussion.

The participants were then asked to complete the questionnaire. Completion of the survey was timed and participants were asked to assess the relevance and validity of each question. Additionally, they were asked to make assessments about the usability of the survey itself and report any complications or need for clarity surrounding the instructions within the survey itself. Participants provided suggestions for improving the clarity and ease of use of the questionnaire in the focus group interview with the researcher. Participants provided all comments and suggestions on a voluntary basis.

At the end of each focus group interview, participants were asked for any additional information and/or perspective. Each focus group interview ended with the researcher thanking the participants and the participants were encouraged to contact the researcher if additional information and/or perspective could be added. Every participant received a thank you note shortly after the interview.

The focus group methodology produces a breadth of information as well as concentrated data on this specific area of interest (Krueger, 1988). The themes from the focus group interviews were then explored more systematically. The information gathered in the focus groups guided the revision of the survey for future participants who were not part of the focus group interviews. The final questionnaire was revised, as a result of participant input.

Instrumentation

Survey

An electronic survey instrument, designed specifically for this study, was used to capture responses from participants. The questionnaire was a 20-item web-based instrument designed to elicit basic demographic information, as well as information pertaining to factors that may have influenced peer ratings in a MRF assessment. A web-based questionnaire was selected as the method to collect this data because it provides a convenient, fast, and cost effective way to reach a large and diverse number of participants. A combination of five-point Likert rating scales, open-ended questions, and demographic questions were used to collect participants' responses and provide consistent, valid and reliable data.

Specific to this research study, the benefit to using a web-based survey is that the survey was designed to change and develop as the respondents answer questions (Pitkow & Recker, 1994). To ensure that participants met the research study requirements, after submitting responses to the qualifying questions, respondents were directed to the next set of questions that were reflective of their response. Specifically, the first question asked respondents to consent to voluntary participation, confirm that they had participated in a MRF assessment as peer rater within the previous 18 months, and were over the age of 18. To be eligible to participate in the study, participants must have responded to these questions appropriately. Pending their response to these questions, the survey prompted the respondent to exit the survey or directed the respondent to the next set of questions to continue with the survey.

An additional benefit to using a web-based survey is that this survey was designed to allow only one response per computer. Additionally, a cutoff date and time that the survey would stop accepting responses had been set. Specific survey parameters that had also been set were that respondents could go back to previous pages in the survey and update existing responses until the survey was finished or until they had exited the survey. Once the survey was finished, the respondent was not be able to re-enter the survey. To preserve anonymity and insure privacy, respondents' Internet Protocol (IP) addresses, the numerical identification assigned to all information technology devices, were not be stored in the survey results.

Five-point Likert Rating Scales

In this questionnaire, five-point Likert rating scales were used to ask participants to indicate the degree to which each of the variables listed may have influenced their assessment of a peer (i.e. co-worker). These rating scale questions required participants to rate an item along a well-defined continuum with five clear choices. Each of the variables listed described potential influences that might have lead to a purposeful adjustment or unintentional bias in their ratings.

Open Ended Questions

Open-ended question asked for comments to provide respondents with the opportunity to add additional information. These comments were summarized and provided more meaning and clarity to the survey results. This use of open-ended questions explored the qualitative, in-depth aspects of respondents' quantitative responses. It gave participants the chance to respond in much greater detail. The option of

"Other" was added as an answer to specific questions and the use of a comments field was included in case participants indicated that they did not find an option that best suited their intended response.

Demographic Questions

Demographic questions were used to identify characteristics such as age, gender, and so forth. Specifically, the demographic questions helped to classify differential or similarities that existed between raters and ratees.

Survey Participants

For the purpose of describing the individuals that participated in the study, the words participants, subjects, and/or respondents will continue to be used. The participants represented a variety of industries and organizations as peer raters in a multirater feedback assessment. These individuals included approximately 62 members of the MN Chapter of the International Society of Performance Improvement (ISPI), nearly 100 members of the Front Range Chapter of the International Society of Performance Improvement (ISPI), roughly 30 members of the Seattle Chapter of the International Society of Performance Improvement (ISPI), nearly 350 members of the Puget Sound Chapter of the American Society of Training and Development (ASTD), and approximately 125 members of the Pacific Northwest Chapter of Organizational Development Network (ODN). These organizations were selected because they offer large sample sizes of raters, contain raters from a variety of industries and organizations, and a range of MRF assessment processes and tools.

Identification of Participants

Prior to the inception of this research, the participants must have completed a formal MRF assessment as peer raters within their organizations of affiliation. The MRF assessment was part of a process sponsored solely by their affiliated organizations and not conducted or implemented by the researcher. Participants were selected based on their completion of a MRF assessment as peer raters. Only those participants meeting the above criteria were included as subjects for this study.

Solicitation of Participants

Participants who were members of the MN Chapter of the International Society of Performance Improvement (ISPI), the Seattle Chapter of the International Society of Performance Improvement (ISPI), the Front Range Chapter of the International Society of Performance Improvement (ISPI), the Puget Sound Chapter of the American Society of Training and Development (ASTD), and the Pacific Northwest Chapter of Organizational Development Network (ODN) were contacted by electronic mail. These associations approved access to these populations and agreed to contribute to this study by including a link to the survey in an email to members and/or posted to the association’s website. Initial contact to participants regarding this study included a letter of invitation sent via electronic mail by the President or an Administrator of the organization. An invitation was also posted on each affiliated website where members were invited to participate in the study.

All participants were asked to voluntarily participate in this study that examined nonperformance factors that may have influenced their assigned ratings to peers in a

MRF assessment. Individuals were informed that their responses would be kept confidential and that information provided would be seen by the researcher only; at no time would individual responses be divulged, nor would any respondents or organizations be identified. Participants were asked to consider the ratings they provided in a MRF assessment of a peer (i.e. co-worker) with which they had the clearest recollection. In order to ensure the confidentiality of raters, ratees, and organizations, participants were not asked or required to identify themselves, specific individuals, and/or the organization that sponsored the MRF assessment.

An advantage to the use of an electronic questionnaire is that it can be delivered to recipients in seconds, rather than in days as with traditional mail. Research shows that response rates are higher with electronic surveys than with paper surveys or interviews. More importantly, research shows that respondents may answer more honestly with electronic surveys than with paper surveys or interviews. The questionnaire was distributed as a link to a web-based survey embedded within an invitation to recipients to participate in the study. This invitation was sent to respondents via an electronic mail invitation and/or posted to the organization’s website.

Participant Incentive

Frick, Bäechtinger, and Reips (1999) conducted an experiment on the effect of incentives on response. They concluded that the chance to win prizes in a lottery resulted in lower dropout rates than in those conditions where no prize drawing entry was offered as an incentive. Based on previous research on response rates, in exchange for their participation, participants’ of this study could elect to be included in a drawing for one of four $25 Amazon gift certificates. Participants had the option to enter into the drawing while keeping their responses anonymous. Upon completion of the survey, participants had two options to select: “Click here to register” in the drawing, or “Done” to exit the survey.

The first URL was the original survey, which contained the actual survey questions that addressed the research questions. Once the participant reached the final page of the survey, if the “Click here to register” link was selected, the respondent was redirected to a second survey containing registration questions giving them the option to enter into the drawing. This second URL was a follow-up survey that asked essential demographic questions, which were kept separate from the original data. In this second survey, demographic questions were asked using open-ended questions to allow participants to enter his/her email, name, and mailing address to register for the drawing. This second survey was not attached in any way to the first. Each survey had its own unique link, which was transparent to participants.

As described, a hyperlink for the second survey was simply embedded on the last page of the first survey with no other survey questions on the page. This was to ensure that all pages of the survey were completed and saved before the participant exited or accessed the second survey to register in the drawing. Participants could select the “Done” button to choose not to be entered, upon which the survey ended.

Anonymity

For the purpose of maintaining anonymity, no information on the first survey was traceable or individually identifiable to any respondent. As a result, all participants were asked permission to participate in the study by his or her implied consent. An opening statement explained the purpose of the study and how the survey information would be used. If participants chose not to participate, they could simply choose not to respond to the survey.

An introduction at the beginning of the questionnaire requested subjects' participation in answering questions about their experiences as peer raters in a multirater feedback process with which they had the clearest recollection. Instructions indicated clearly that individual rater responses to the research questionnaire would not be shared with either the ratee or the organization with which participants or ratees were affiliated. Participants were notified in writing that the entire dissertation would be published and the results of the survey would be used for research purposes only. This final manuscript does not include any actual names of participants or other personal information such as addresses, e-mail addresses, or telephone numbers. Individual survey responses and registration information were kept in strict confidence.

As described, participants received access to the survey using a URL link sent via electronic email or as accessed on the website of the corresponding association. The questionnaire and the data collected was hosted and stored in a confidential electronic database. Participants' responses did not identify the participant, the ratee, or the organization and all responses were kept confidential. Data was stored on a password protected personal computer in the home office of the researcher. All coded documentation and completed survey responses were secured by the researcher and will be retained for five years for future research.

Prior to administering the survey, the researcher tested the survey for reliability and validity. Capella University's Institutional Review Board (IRB) approval was granted prior to administration of this study.

Measures

In order to examine the hypotheses delineated in this research, the questionnaire was designed to solicit participant demographic information and potential factors which may have influenced their assigned ratings. Presented first in the questionnaire was a series of questions designed to elicit basic demographic information from participants. Variables based on basic demographic information included: (a) gender of rater and ratee, and (b) age differential between the rater and ratee. In additional to standard questions (i.e. gender and age), other factors of particular interest in this study include: (c) tenure, (d) concern for anonymity / confidentiality, (e) comfort with process, (f) opportunity to observe, (g) friendship, (h) likeability, (i) competition, (j) acquaintanceship, (k) training, (l) purpose of assessment, (m) rater selection process, (n) favorability of the overall rating, and (o) task association. These variables were compared to raters' perception of accuracy.

Dependent Variables

The dependent variable was the rater's honest use of the performance rating scales based solely on performance.

1. Perception of Accuracy. Raters were asked to indicate their level of agreement to a statement about their confidence in the accuracy of the ratings they assigned to the ratee compared to actual performance. Options varied on a five-level Likert scale: a) Strongly Agree, b) Agree, c) Neither Agree nor Disagree, d) Disagree, or e) Strongly Disagree.

Independent Variables

The following were used as the independent variables:

1. Gender of Rater / Ratee. Gender of the rater and of the ratee were measured on a multiple choice scale as: a) male or b) female. The demographic variable of gender has been the subject of many performance appraisal studies (Hartel, Douthitt, Hartel, & Douthitt, 1999; Landy & Farr, 1980; Bigoness, 1976; Hamner, Kim, Baird, & Bigoness, 1974; Pulakos & Wexley, 1983; Nevill, Stephenson, & Philbrick, 1983). There was sufficient evidence, from these studies, to believe that gender would have an effect on performance ratings.

2. Age of Rater / Ratee. The approximate age of the rater was measured on a multiple-choice scale: a) 18-25, b) 26-35, c) 36-45, d) 46-55, or e) 56 or over. Similarly, the age differential between the ratee / rater was also measured on a multiple-choice scale: a) I am younger, b) We are about the same age, c) I am older, or d) I do not know. The age variable was included in this study to examine whether maturity level or age differential has an impact on performance ratings by peers.

3. Purpose of Assessment. The purpose of a performance appraisal can be classified as either administrative, developmental, or combination. Evaluations used to determine a component of an employee’s pay are classified as administrative. Evaluations used for strictly developmental purposes are classified as developmental. If the purpose includes both administrative and developmental objectives, the purpose is classified as a combination. Farh, Cannella, and Bedeian (1991) used a quasi-experimental design to determine the effects of purpose (administrative/evaluative versus developmental) on peer rating quality. Farh et al. (1991) reported that purpose of peer ratings had a significant impact on the quality of peer ratings. Peer ratings for evaluative purposes had greater leniency, greater halo effect, more uniformity, and less inter-rater reliability than peer evaluations conducted for developmental purposes. Raters were asked to identify the purpose of the performance evaluation. The multiple choice options were: a) administrative, b) developmental, c) combination, d) I do not know, or e) other.

4. Rater Selection Process. Most companies allow individuals to choose their own raters. The natural inclination is for ratees to select people they like and who they think like them in return. People are inclined to choose work friends because they want to be perceived positively and to receive favorable ratings and they think that having friends rate them will help accomplish this. However, friends can be brutally honest, especially when assured anonymity. Raters were asked to identify the rater selection process. The multiple choice options were: a) selected by the organization, b) selected by the ratee, c) selected by lottery/random, d) I do not know, or e) other.

5. Familiarity/Acquaintanceship. It may seem reasonable to assume that the “accuracy” of a rating corresponds with, or at least is related to, how well and/or how long the rater knows the ratee. Research has suggested that the more familiar the rater is with the ratee the higher the evaluation ratings will be. As a measure for rater / ratee familiarity, raters were asked to identify the amount of time that they were a peer of the ratee on a multiple-choice scale: a) 1 year or less, b) 1 year + but less than 3, c) 3 years + but less than 5, d) 5 years + but less than 10, and e) 10 + years.

6. Tenure. The duration the rater had been with the organization may be an important variable to indicate job experience and as a measure for organizational familiarity. The approximate tenure of the rater was measured based on longevity with the organization on a multiple-choice scale: a) 1 year or less, b) 1 year + but less than 3, c) 3 years + but less than 5, d) 5 years + but less than 10, and e) 10 + years.

7. Concern for Anonymity/Confidentiality. The level of concern for anonymity and confidentiality captured the perceived privacy of the rater in the peer evaluation. Raters were asked to indicate their level of agreement to a statement about their belief that anonymity was protected while assessing the ratee on a five-level Likert scale: a) Strongly agree, b) Agree, c) Neither Agree nor Disagree, d) Disagree, or e) Strongly Disagree.

8. Comfort with Process. Additionally, raters were asked to indicate their level of agreement to a statement about their level of comfort with the MRF process on the same five-level Likert scale.

9. Opportunity to Observe. Some data sources are more valid than others because of their observational skills and their opportunity to observe actual performance. Ideally, all raters should have sufficient opportunities to observe the ratee in work situations. Raters were asked to indicate their level of agreement to a statement about the opportunity they had to observe the ratee. Options varied on a five-level Likert scale: a) Strongly Agree, b) Agree, c) Neither Agree nor Disagree, d) Disagree, or e) Strongly Disagree.

10. Likeability and Friendship. One aspect of MRF that researchers have largely overlooked is the possibility that the relationship between rater and ratee may influence ratings. This oversight is cause for some concern, as MRF assessments depend on the quality of ratings from multiple sources (Antonioni & Park, 2001). Raters were asked to identify their level of agreement to a statement about their friendship with the ratee as well as their identification of the ratee as a “likeable” person. Options varied on a five-level Likert scale: a) Strongly Agree, b) Agree, c) Neither Agree nor Disagree, d) Disagree, or e) Strongly Disagree.

11. Competition. Raters were asked to identify their level of agreement to a statement about the degree of competition that they believed existed between themselves and the ratee. Options varied on the same five-level Likert scale.

12. Training Received. Rater training might focus on the rating scale of the MRF assessment, explanations of the content of the rating instrument, and clarification of the purpose of the assessment. Raters were asked to identify their level of agreement to a statement about the effectiveness of the training they received to prepare them for the MRF assessment. Options varied on a five-level Likert scale: a) Strongly Agree, b) Agree, c) Neither Agree nor Disagree, d) Disagree, or e) Strongly Disagree.

13. Task Familiarity. Raters were asked to identify their level of agreement to a statement about how familiar they were with the assigned responsibilities and tasks assigned to the ratee. Options varied on the same five-level Likert scale.

14. Assigned Rating. Raters were asked to identify their level of agreement to a statement about the assigned rating or overall assessment of the ratee. When asked if the overall assessment could be described as “favorable” or “positive”, participants were given options on a five-level Likert scale: a) Strongly Agree, b) Agree, c) Neither Agree nor Disagree, d) Disagree, or e) Strongly Disagree.

15. Potential Bias. Raters were asked to identify their level of agreement to a statement about the influence of factors, other than performance, that may have influenced the ratings they assigned to the ratee. Options varied on the same five-level Likert scale.

Procedures

In order to establish a wide foundation for the research questions, a diverse group of participants was used in this research study. A focus on a single industry or organization may have limited the study or hindered participants from being truthful. The participants selected for this study provided data from a broad spectrum of respondents; from different industries, from corporate to non-profit, and different layers of management in multiple organizations and organization types.

The participants of this study were asked first to think of a peer (i.e. co-worker) for whom they completed a MRF assessment. They were subsequently asked to respond to a series of questions with this assessment in mind.

Data Collection

The data collection process was conducted over a two-week timeframe. Once the individual data was collected, the results were downloaded from the online database to a computer software program, SPSS version 15.0. The researcher generated reports from the SPSS software database to analyze the results. The final data will be stored for five years to allow future research to continue to add to the existing database. Raw data will be stored for up to three years and then discarded.

Data Analysis

This section will outline the steps that were taken to analyze the data after collection. Descriptive statistics were computed for the demographic data and the results were included in the analysis portion of this study. Each of the hypotheses were tested using a Chi Square Test of Independence in SPSS version 15.0. For this analysis, a relationship is tested between two nominal or ordinal variables. It is a perfect test to use with ordinal variables using the Likert scale (Mirabella, 2006). The Chi Square Test of Independence was used to compare the questions relating to the nonperformance factors associated with the ratee to the questions relating to the overall rating assigned to the ratee’s performance.

Ethical Considerations

Researchers must be aware of ethical considerations inherent in their studies and must take responsibility to ensure that their studies will do no harm and pose no risk to the participants. Researchers must take measures to protect the rights and welfare of the study participants. This research study was granted approval by Capella University’s

Institutional Review Board (IRB) and followed all guidelines for research involving human subjects to ensure that the study met all ethical considerations.

For example, the purpose of the study was fully disclosed to participants. Participants were also assured that the results of their feedback would be reported in aggregate form and all data would remain anonymous in that there will be no way of identifying participants in the study. To foster anonymity, the names of participants, email addresses, and/or contact information of the entire participant population were never known by the researcher. All respondents participated voluntarily and were allowed to terminate involvement in the study at any time by simply not responding to the invitation or by exiting the web-page containing the survey.

Additionally, Secure Sockets Layer (SSL) was used to encrypt the survey. Secure Sockets Layer (SSL) is used for transmitting information privately over the Internet. This prevented participants' IP addresses from being stored in the survey results. Participants were also provided with the contact information of the researcher, the Institutional Review Board (IRB), and the dissertation chairperson for questions or concerns about the study.

Conclusion

Chapter 3 discussed the methodology that was used to conduct the research. This chapter included a discussion of the research design, participants involved in the study, maintenance of anonymity, procedures used to conduct the study, instruments used to collect the data, and an overview of the data collection process.

Multirater feedback continues to emerge as a popular performance appraisal method. This method promises performance evaluation data that is less biased, more reliable, and more valid that the traditional supervisor-only appraisal method. However, given that it is a common practice in a MRF process for ratees to select the raters of their performance, and there may be nonperformance factors that influence peer raters, the method may actually be introducing new biases. Given this, this study conducted a survey to assess the effects that individual characteristics and nonperformance variables may have on peer rater feedback.

CHAPTER 4. DATA COLLECTION AND ANALYSIS

Data Collected

The purpose of this study was to identify variables that may affect peer raters' responses to a multirater feedback instrument. Numerous variables anticipated to affect the peer rater's response to an MRF instrument have been examined. Specifically, this study examined the relationship between raters' perception of the accuracy of their ratings to nonperformance variables such as: (a) gender, (b) age, (c) tenure, (d) concern for anonymity / confidentiality, (e) comfort with process, (f) opportunity to observe, (g) training, (h) purpose of assessment, (i) task association, (j) friendship, (k) likeability, (l) competition, (m) acquaintanceship, (n) favorability of the overall rating, and (o) the rater selection process.

This study answered a series of four research questions through the development of nineteen relevant hypotheses and use of statistical techniques to either support or not support them. This chapter reports the data analysis and results of the study. The results of the hypotheses testing are introduced followed by an analysis of these results. Since all of the variables tested were categorical in nature, the Chi Square Test of Independence was conducted to determine the statistical significance of each hypothesis. A significance level of .05 was used for each hypothesis test.

Participants and Procedures

Approximately 667 potential participants were included in this study. These subjects represented a variety of industries and organizations as peer raters in a multirater feedback assessment. They were made up of approximately 62 members of the MN

Chapter of the International Society of Performance Improvement (ISPI), nearly 100 members of the Front Range Chapter of the International Society of Performance Improvement (ISPI), roughly 30 members of the Seattle Chapter of the International Society of Performance Improvement (ISPI), nearly 350 members of the Puget Sound Chapter of the American Society of Training and Development (ASTD), and approximately 125 members of the Pacific Northwest Chapter of Organizational Development Network (ODN). These organizations were selected because they offer large sample sizes of raters and contain raters using a variety of MRF assessment processes and tools.

Subjects were invited to participate in the study via an email message which gave a brief overview of the purpose of the study, outlined confidentiality information, and also contained instructions for the online questionnaire. Of the potential participants, 136 individuals responded. Out of the 136 responses, 20 were disregarded due to incomplete answers on the survey and/or did not meet the criteria. A total of 116 participant responses were included in this study, which represents over 16 percent of the population tested. Response rates for a web-based survey are expected to be low, especially when invitations are sent to e-mails addresses. For example, it is expected that a portion of these address are invalid due, in part, to employees changing jobs, out dated e-mail addresses, or to blocked access to messages with links or attachments. There was no reason to suspect participants to have a fear of reprisal, thus impacting response bias. It is assumed that the results are representative of the targeted population as the association membership lists include professionals from diverse companies varying in size and industry.

Of the responses, there were 73 female participants and 43 male participants. The age of the participants ranged from 18-56 + years. To participate in the study, all subjects must have completed a formal MRF assessment as peer raters within the previous 18 months. A MRF assessment was not conducted or implemented by the researcher; the process was sponsored solely by participants' affiliated organization. Participants were selected based on their completion of a MRF assessment as peer raters. Only those participants meeting the above criteria were included as subjects for this study.

Instruments and Data

Data was gathered using a mixed methodology. The first tool was a focus group interview and the second instrument was an online survey. Two focus groups were led by the researcher. Focus group participants were asked to complete a proposed questionnaire. Completion of the survey was timed and participants were asked to assess the relevance and validity of each question. Additionally, they were asked to make assessments about the usability of the survey, itself, and report any complications or need for clarity surrounding the instructions contained within the survey. In the focus group interview with the researcher, participants provided suggestions for improving the clarity and usability of the questionnaire. Participants provided all comments and suggestions on a voluntary basis.

Each focus group session began with the researcher welcoming the participants and thanking them for their time. Next, the researcher briefly described the purpose of the focus group and the research project, and then explained that the interview would last approximately 45-60 minutes, that it would be recorded for transcription purposes only, and that all names and data would be kept confidential. The researcher reviewed a description of the role as facilitator, the role of the participants, and the ground rules for participation. The researcher communicated that all members of the group should be allowed to participate equally and that only one person should speak at a time.

The information gathered in the focus groups guided the revision of the survey for future participants who were not part of the focus group interviews. The results from the focus groups served as a source of data for the revision of the survey questionnaire used in the quantitative aspect of this research.

The electronic survey instrument used data provided from the focus groups, specifically to capture responses from participants. The questionnaire was a 20-item web-based instrument designed to elicit basic demographic information, as well as information pertaining to factors that may have influenced peer ratings in a MRF assessment. A web-based questionnaire served as the ideal method to collect this data because it provides a convenient, fast, and cost effective way to reach a large and diverse number of participants. A combination of five-point Likert rating scales, open-ended questions, and demographic questions collected participants’ responses to provide consistent, valid, and reliable data.

This study addressed four research questions through the development of relevant hypotheses. Statistical techniques were used to either support or not support a total of nineteen hypotheses. The following research questions have been examined in this study:

Research Question 1: What is the relationship between the accuracy of multirater feedback ratings from peers versus the nonperformance factors / demographics of the rater and ratee?

Thirteen hypotheses were developed to support this question in determining the existence of a relationship between the accuracy of ratings from peers versus nonperformance factors or demographics of the rater and ratee. They represented: (a) gender, (b) age, (c) tenure, (d) concern for anonymity / confidentiality, (e) comfort with process, (f) opportunity to observe, (g) training, (h) purpose of assessment, (i) task association/familiarity, and (j) other potential nonperformance factors, and were evaluated against the constant variable of rating accuracy.

Hypothesis 1a: Accuracy of multirater feedback ratings from peers is independent of the gender of the rater.

This hypothesis was evaluated by comparing responses to question 2, “My gender is:” and question 19 on the web-based survey. Question 2 inquired of the gender of the rater, whereas question 19, “I am confident that I assigned accurate ratings to this individual and provided an accurate assessment of their performance.” explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1b: Accuracy of multirater feedback ratings from peers is independent of the gender of the ratee.

This hypothesis was evaluated by comparing responses to question 3, “The gender of the ratee is:” and question 19 on the web-based survey. Question 3 inquired of the gender of the ratee, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1c: Accuracy of multirater feedback ratings from peers is independent of whether the rater and ratee are the same gender.

This hypothesis was evaluated by comparing responses to question 2, “The gender of the ratee is:”, with question 3, “The gender of the ratee is:” and question 19 on the web-based survey. Question 2 and 3 inquired of the gender of the rater and the ratee, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1d: Accuracy of multirater feedback ratings from peers is independent of the age of the rater.

This hypothesis was evaluated by comparing responses to question 4, “My age is:” and question 19 on the web-based survey. Question 4 inquired of the age of the rater, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1e: Accuracy of multirater feedback ratings from peers is independent of the difference in age between the rater and ratee.

This hypothesis was evaluated by comparing responses to question 5, “The age differential between the ratee and me:” and question 19 on the web-based survey. Question 5 inquired of the difference in age between the rater and ratee, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1f: Accuracy of multirater feedback ratings from peers is independent of the tenure of the rater with the organization.

This hypothesis was evaluated by comparing responses to question 9, “I was employed by the organization where the evaluation was conducted for a total of:” and question 19 on the web-based survey. Question 9 inquired of the tenure of the rater with the organization at the time of the peer assessment, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1g: Accuracy of multirater feedback ratings from peers is independent of the level of concern for anonymity and confidentiality of the rater.

This hypothesis was evaluated by comparing responses to question 10, “I believe that my evaluation of this individual was confidential and the feedback I provided was not individually identifiable, disclosed, or known by the ratee” and question 19 on the web-based survey. Question 10 inquired of the level of concern for anonymity and confidentiality of the rater, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1h: Accuracy of multirater feedback ratings from peers is independent of the raters’ comfort level with the multirater feedback process.

This hypothesis was evaluated by comparing responses to question 11, “I was comfortable with the process of providing feedback to this individual” and question 19 on the web-based survey. Question 11 inquired of the raters’ comfort level with the multirater feedback process, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1i: Accuracy of multirater feedback ratings from peers is independent of the opportunity raters had to observe the ratee.

This hypothesis was evaluated by comparing responses to question 12, “I had sufficient opportunity to observe the aspects of the individual's performance that I was asked to evaluate.” and question 19 on the web-based survey. Question 12 inquired of the opportunity raters had to observe the ratee, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1j: Accuracy of multirater feedback ratings from peers is independent of the influence of rater training.

This hypothesis was evaluated by comparing responses to question 16, “The training and/or instruction I received (in regards to the evaluation process and use of the rating scales) prior to completing the evaluation was sufficient/effective/informative.” and question 19 on the web-based survey. Question 16 inquired of the influence of rater training, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1k: Accuracy of multirater feedback ratings from peers is independent of the purpose of peer ratings.

This hypothesis was evaluated by comparing responses to question 6, “The organizational purpose of the peer evaluation was:” and question 19 on the web-based survey. Question 6 inquired of the purpose of peer ratings, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1l: Accuracy of multirater feedback ratings from peers is independent of the familiarity of peer raters to the assigned responsibilities and tasks of the ratee.

This hypothesis was evaluated by comparing responses to question 17, “I was familiar with the responsibilities and tasks assigned to this individual.” and question 19 on the web-based survey. Question 17 inquired of the familiarity of peer raters to the assigned responsibilities and tasks of the ratee, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 1m: Accuracy of multirater feedback ratings from peers is independent of nonperformance factors.

This hypothesis was evaluated by comparing responses to question 20, “Factors other than performance may have influenced one or more of the ratings I assigned to this individual.” and question 19 on the web-based survey. Question 20 inquired of the influence of factors other than performance, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Research Question 2: What is the relationship between the accuracy of multirater feedback ratings from peers versus the personal relationship between the rater and ratee?

To support this question in determining the existence of a relationship between the accuracy of ratings from peers versus the personal relationship between the rater and ratee, four hypotheses were developed representing: (a) friendship, (b) likeability, (c) competition, and (c) how long the rater has known the ratee, and were evaluated against the constant variable of rating accuracy.

Hypothesis 2a: Accuracy of multirater feedback ratings from peers is independent of the friendship between rater and ratee.

This hypothesis was evaluated by comparing responses to question 13, “In addition to my professional relationship, I would describe this individual as a ‘friend’.” and question 19 on the web-based survey. Question 13 inquired of the influence of the friendship between rater and ratee, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 2b: Accuracy of multirater feedback ratings from peers is independent of how well liked the ratee is by the rater.

This hypothesis was evaluated by comparing responses to question 14, “I would describe this individual as a ‘likeable’ person.” and question 19 on the web-based survey. Question 14 inquired of the influence of how well liked the ratee is by the rater, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 2c: Accuracy of multirater feedback ratings from peers is independent of the degree of competition that exists between the rater and the ratee.

This hypothesis was evaluated by comparing responses to question 15, “I would describe this individual as my ‘competition’ (for pay, awards, recognition, promotion, etc.).” and question 19 on the web-based survey. Question 15 inquired of the influence of the degree of competition that exists between the rater and the ratee, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Hypothesis 2d: Accuracy of multirater feedback ratings from peers is independent of how long the rater has known the ratee.

This hypothesis was evaluated by comparing responses to question 8, “I was a peer of the ratee for:” and question 19 on the web-based survey. Question 8 inquired of how long the rater had known the ratee, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Research Question 3: What is the relationship between the accuracy of multirater feedback ratings from peers versus the selection process for peer raters?

To support this question in determining the existence of a relationship between the accuracy of ratings from peers versus the selection process for peer raters, one hypothesis was developed representing the rater selection process and evaluated against the constant variable of rating accuracy.

Hypothesis 3: Accuracy of multirater feedback ratings from peers is independent of the rater selection process.

This hypothesis was evaluated by comparing responses to question 7, “I was assigned/selected to provide feedback as a peer rater by:” and question 19 on the web-based survey. Question 7 inquired of the rater selection process, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Research Question 4: What is the relationship between the accuracy of multirater feedback ratings from peers as opposed to the favorability of the overall rating?

To support this question in answering the existence of a relationship between the accuracy of ratings from peers versus the favorability of the overall score, one hypothesis was developed representing the favorability of the overall rating and evaluated against the constant variable of rating accuracy.

Hypothesis 4: Accuracy of multirater feedback ratings from peers is independent of the favorability of the overall rating.

This hypothesis was evaluated by comparing responses to question 18, “The overall evaluation/rating/score I provided for this individual could be described as "favorable" or positive.” and question 19 on the web-based survey. Question 18 inquired of the influence of the favorability of the overall rating, whereas question 19 explored the accuracy of the multirater feedback ratings, as reported by peer raters.

Results

The data collected through the online survey was imported into SPSS for quantitative analysis. The Chi Square Test of Independence was conducted to determine the statistical significance of each hypothesis. In this research study, a minimum significance level of .05 was used for each test. This means that the differences will be statistically significant if the results would have occurred by chance less than 5 times out of 100. It is reported as $p < .05$. When the statistical difference is strong, the p value will be reported as $p <.01$, which means that the results would have occurred by chance less than 1 time in 100. If there is no significant difference, the actual p value will be reported.

The following hypotheses related to research question 1 have been tested:

Hypothesis 1a: Accuracy of multirater feedback ratings from peers is independent of the gender of the rater.

Of the responses, there were 73 female participants and 43 male participants.

Table 2. Cross tabulation for H1a: Gender of the Rater

	Accurate		
	Positive	Negative	Total
Male	17	26	43
Female	29	44	73
Total	46	70	116

Table 3. Chi-Square Test for H1a: Gender of the Rater

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.000	1	.984

For this and all hypotheses, “positive” refers to responses in the highest category only (i.e., “strongly agree), while “negative” refers to all other responses. This grouping is based on the premise that any response less than the highest category is indicative of doubt on the part of the respondent, and this study is about accuracy and integrity of ratings. With a p-value of .984, which is greater than .05, the hypothesis is not rejected. Therefore, there is insufficient evidence to conclude that the accuracy of feedback ratings depends on the gender of the rater.

Hypothesis 1b: Accuracy of multirater feedback ratings from peers is independent of the gender of the ratee.

Table 4. Cross tabulation for H1b: Gender of the Ratee

	Accurate		
	Positive	Negative	Total
Male	19	31	50
Female	27	39	66
Total	46	70	116

Table 5. Chi-Square Test for H1b: Gender of the Ratee

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.101	1	.751

With a p-value of .751, which is greater than .05, the hypothesis is not rejected. In this case, there is insufficient evidence to conclude that the accuracy of feedback ratings depends on the gender of the ratee.

Hypothesis 1c: Accuracy of multirater feedback ratings from peers is independent of whether the rater and ratee are the same gender.

Table 6. Cross tabulation for H1c: Gender Difference

	Accurate		
	Positive	Negative	Total
Different genders	12	21	33
Same genders	34	49	83
Total	46	70	116

Table 7. Chi-Square Test for H1c: Gender Difference

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.209	1	.648

With a p-value of .648, which is greater than .05, the hypothesis is not rejected. As a result, there is insufficient evidence to conclude that the accuracy of feedback ratings depends on the gender difference between the rater and ratee.

Hypothesis 1d: Accuracy of multirater feedback ratings from peers is independent of the age of the rater.

The age categories within this query were defined as 18-35 years, 36-45 years, 46-55 years, and over 55 years. The youngest group of respondents, ranging in age from 18-35 represented 22% of the sample, while 27% was made up of the second youngest of the group, with an age range of 36-45. The category of 46-55 represented 33% of the sample, while the oldest category, over 55, contributed 18% of the sample population.

Table 8. Cross tabulation for H1d: Age of the Rater

	Accurate		
	Positive	Negative	Total
18-35	6	20	26
36-45	12	19	31
46-55	17	21	38
Over 55	11	10	21
Total	46	70	116

Table 9. Chi-Square Test for H1d: Age of the Rater

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.829	3	.185

With a p-value of .185, which is greater than .05, the hypothesis is not rejected. Consequently, there is insufficient evidence to conclude that the accuracy of feedback ratings depends on the age of the rater.

Hypothesis 1e: Accuracy of multirater feedback ratings from peers is independent of the difference in age between the rater and ratee.

The categories within this question were defined as: younger, about the same age, and older.

Table 10. Cross tabulation for H1e: Age Difference

	Accurate		
	Positive	Negative	Total
Rater younger	9	11	20
About the same age	17	36	53
Rater older	20	23	43
Total	46	70	116

Table 11. Chi-Square Test for H1e: Age Difference

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.356	2	.308

With a p-value of .308, which is greater than .05, the hypothesis is not rejected. Thus, there is insufficient evidence to conclude that the accuracy of feedback ratings depends on the difference in age between the rater and ratee.

Hypothesis 1f: Accuracy of multirater feedback ratings from peers is independent of the tenure of the rater with the organization.

The categories within this query were defined as 1 year or less, 1 - 3 years, 3 - 5 years, 5 - 10 years, and more than 10 years.

Table 12. Cross tabulation for H1f: Tenure of the Rater

	Accurate		
	Positive	Negative	Total
1 year or less	7	7	14
1 - 3 years	10	22	32
3 - 5 years	9	12	21
5 - 10 years	12	11	23
More than 10 years	8	18	26
Total	46	70	116

Table 13. Chi-Square Test for H1f: Tenure of the Rater

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.025	4	.403

With a p-value of .403, which is greater than .05, the hypothesis is not rejected. As a result, there is insufficient evidence to conclude that the accuracy of feedback ratings depends on the tenure of the rater with the organization.

Hypothesis 1g: Accuracy of multirater feedback ratings from peers is independent of the level of concern for anonymity and confidentiality of the rater.

Since a ‘strongly agree’ response is the only response that equates to a fully committed answer, the responses were coded into two possible categories: ‘strongly agree’ (coded as “positive”) and ‘less than strongly agree’ (coded as “negative”).

Table 14. Cross tabulation for H1g: Confidentiality

	Accurate		
	Positive	Negative	Total
Positive	22	17	39
Negative	24	53	77
Total	46	70	116

Table 15. Chi-Square Test for H1g: Confidentiality

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.893	1	.009

With a p-value of .009, which is less than .05, the hypothesis is rejected. It can be concluded that the accuracy of feedback ratings depends on the level of concern for anonymity and confidentiality of the rater. Raters who are less concerned about their anonymity and confidentiality were more likely to identify their ratings as “accurate.” This means that some raters attribute the inaccuracy of their ratings to their concern for anonymity and confidentiality.

Hypothesis 1h: Accuracy of multirater feedback ratings from peers is independent of the raters' comfort level with the multirater feedback process.

Since a 'strongly agree' response is the only response that equates to a fully committed answer, the responses were coded into two possible categories: 'strongly agree' (coded as "positive") and 'less than strongly agree' (coded as "negative").

Table 16. Cross tabulation for H1h: Comfort Level

	Accurate		
	Positive	Negative	Total
Positive	34	23	57
Negative	12	47	59
Total	46	70	116

Table 17. Chi-Square Test for H1h: Comfort Level

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	18.721	1	.000

With a p-value of .000, which is less than .05, the hypothesis is rejected. It can be concluded that the accuracy of feedback ratings depends on the raters' comfort level with the multirater feedback process. Raters more comfortable with the multirater feedback process were more likely to identify their ratings as "accurate." This means that some raters attribute the inaccuracy of their ratings to their lack of comfort with the multirater feedback process.

Hypothesis 1i: Accuracy of multirater feedback ratings from peers is independent of the opportunity raters had to observe the ratee.

Since a ‘strongly agree’ response is the only response that equates to a fully committed answer, the responses were coded into two possible categories: ‘strongly agree’ (coded as “positive”) and ‘less than strongly agree’ (coded as “negative”).

Table 18. Cross tabulation for H1i: Opportunity to Observe

	Accurate		
	Positive	Negative	Total
Positive	38	19	57
Negative	8	51	59
Total	46	70	116

Table 19. Chi-Square Test for H1i: Opportunity to Observe

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	34.169	1	.000

With a p-value of .000, which is less than .05, the hypothesis is rejected. It can be concluded that the accuracy of feedback ratings depends on the opportunity raters had to observe the ratee. Raters with greater opportunity to observe the ratee were more likely to identify their ratings as “accurate.” This means that some raters attribute the inaccuracy of their ratings to their lack of opportunity to observe the ratee.

Hypothesis 1j: Accuracy of multirater feedback ratings from peers is independent of the influence of rater training.

Since a ‘strongly agree’ response is the only response that equates to a fully committed answer, the responses were coded into two possible categories: ‘strongly agree’ (coded as “positive”) and ‘less than strongly agree’ (coded as “negative”).

Table 20. Cross tabulation for H1j: Training

	Accurate		
	Positive	Negative	Total
Positive	14	13	27
Negative	32	57	89
Total	46	70	116

Table 21. Chi-Square Test for H1j: Training

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.188	1	.139

With a p-value of .139, which is greater than .05, the hypothesis is not rejected. As a result, there is insufficient evidence to conclude that the accuracy of feedback ratings depends on the influence of rater training.

Hypothesis 1k: Accuracy of multirater feedback ratings from peers is independent of the purpose of peer ratings.

The categories were defined as: administrative, developmental, combination, and other.

Table 22. Cross tabulation for H1k: Purpose

	Accurate		
	Positive	Negative	Total
Administrative	7	8	15
Developmental	17	33	50
Combination	21	29	50
Other	1	0	1
Total	46	70	116

Table 23. Chi-Square Test for H1k: Purpose

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.613	3	.455

With a p-value of .455, which is greater than .05, the hypothesis is not rejected. Consequently, there is insufficient evidence to conclude that the accuracy of feedback ratings depends on the purpose of peer ratings.

Hypothesis 1l: Accuracy of multirater feedback ratings from peers is independent of the familiarity of peer raters to the assigned responsibilities and tasks of the ratee.

Since a ‘strongly agree’ response is the only response that equates to a fully committed answer, the responses were coded into two possible categories: ‘strongly agree’ (coded as “positive”) and ‘less than strongly agree’ (coded as “negative”).

Table 24. Cross tabulation for H1l: Familiarity of Tasks

	Accurate		
	Positive	Negative	Total
Positive	32	24	56
Negative	14	46	60
Total	46	70	116

Table 25. Chi-Square Test for H1l: Familiarity of Tasks

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	13.836	1	.000

With a p-value of .000, which is less than .05, the hypothesis is rejected. It can be concluded that the accuracy of feedback ratings depends on the familiarity of peer raters to the assigned responsibilities and tasks of the ratee. Raters who are more familiar with the assigned responsibilities and tasks of the ratee were more likely to identify their ratings as “accurate.” This means that some raters attribute the inaccuracy of their ratings to their lack of familiarity with the responsibilities and tasks of the ratee.

Hypothesis 1m: Accuracy of multirater feedback ratings from peers is independent of nonperformance factors.

Respondents were asked to respond to a direct question of whether the performance rating they assigned to the ratee was influenced by factors other than performance. Since a ‘strongly agree’ response is the only response that equates to a fully committed answer, the responses were coded into two possible categories: ‘strongly agree’ (coded as “positive”) and ‘less than strongly agree’ (coded as “negative”).

Table 26. Cross tabulation for H1m: Nonperformance Factors

	Accurate		
	Positive	Negative	Total
Positive	4	4	8
Negative	42	66	108
Total	46	70	116

Table 27. Chi-Square Test for H1m: Nonperformance Factors

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.384	1	.535

With a p-value of .535, which is greater than .05, the hypothesis is not rejected. Therefore, there is insufficient evidence to conclude that the accuracy of feedback ratings was influenced by factors other than performance, as reported by the raters.

The following hypotheses related to research question 2 have been tested:

Hypothesis 2a: Accuracy of multirater feedback ratings from peers is independent of the friendship between rater and ratee.

Since a ‘strongly agree’ response is the only response that equates to a fully committed answer, the responses were coded into two possible categories: ‘strongly agree’ (coded as “positive”) and ‘less than strongly agree’ (coded as “negative”).

Table 28. Cross tabulation for H2a: Friendship

	Accurate		
	Positive	Negative	Total
Positive	9	2	11
Negative	37	68	105
Total	46	70	116

Table 29. Chi-Square Test for H2a: Friendship

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	9.028	1	.003

With a p-value of .003, which is less than .05, the hypothesis is rejected. It can be concluded that the accuracy of feedback ratings depends on the friendship established between rater and ratee. Raters who identified as having a friendship with the ratee were more likely to identify their ratings as “accurate.” This means that some raters attribute the inaccuracy of their ratings to their lack of friendship with the ratee.

Hypothesis 2b: Accuracy of multirater feedback ratings from peers is independent of how well liked the ratee is by the rater.

Since a ‘strongly agree’ response is the only response that equates to a fully committed answer, the responses were coded into two possible categories: ‘strongly agree’ (coded as “positive”) and ‘less than strongly agree’ (coded as “negative”).

Table 30. Cross tabulation for H2b: Likeability

	Accurate		
	Positive	Negative	Total
Positive	14	14	28
Negative	32	56	88
Total	46	70	116

Table 31. Chi-Square Test for H2b: Likeability

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.651	1	.199

With a p-value of .199, which is greater than .05, the hypothesis is not rejected. As a result, there is insufficient evidence to conclude that the accuracy of feedback ratings was influenced by how well liked the ratee is by the rater.

Hypothesis 2c: Accuracy of multirater feedback ratings from peers is independent of the degree of competition that exists between the rater and the ratee.

Since a ‘strongly agree’ response is the only response that equates to a fully committed answer, the responses were coded into two possible categories: ‘strongly agree’ (coded as “positive”) and ‘less than strongly agree’ (coded as “negative”).

Table 32. Cross tabulation for H2c: Competition

	Accurate		
	Positive	Negative	Total
Positive	2	7	9
Negative	44	63	107
Total	46	70	116

Table 33. Chi-Square Test for H2c: Competition

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.239	1	.266

With a p-value of .266, which is greater than .05, the hypothesis is not rejected. Consequently, there is insufficient evidence to conclude that the accuracy of feedback ratings was influenced by the degree of competition that exists between the rater and the ratee.

Hypothesis 2d: Accuracy of multirater feedback ratings from peers is independent of how long the rater has known the ratee.

The categories for this query were defined as: 1 year or less, 1 - 3 years, 3 - 5 years, and more than 5 years.

Table 34. Cross tabulation for H2d: Familiarity with Ratee

	Accurate		
	Positive	Negative	Total
1 year or less	12	16	28
1 - 3 years	15	39	54
3 - 5 years	8	11	19
More than 5 years	11	4	15
Total	46	70	116

Table 35. Chi-Square Test for H2d: Familiarity with Ratee

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10.461	3	.015

With a p-value of .015, which is less than .05, the hypothesis is rejected. It can be concluded that the accuracy of feedback ratings depends on how long the rater has known the ratee. The greater number of years that the rater identified with having known the ratee the more likely they were to identify their ratings as “accurate.” This means that some raters attribute the inaccuracy of their ratings to their unfamiliarity with the ratee.

The following hypothesis related to research question 3 has been tested.

Hypothesis 3: Accuracy of multirater feedback ratings from peers is independent of the rater selection process.

The categories for this query were defined as: organization, ratee, and other.

Table 36. Cross tabulation for H3: Selection Process

	Accurate		
	Positive	Negative	Total
Organization	27	37	64
Ratee	17	31	48
Other	2	2	4
Total	46	70	116

Table 37. Chi-Square Test for H3: Selection Process

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.711	2	.701

With a p-value of .701, which is greater than .05, the hypothesis is not rejected. As a result, there is insufficient evidence to conclude that the accuracy of feedback ratings depends on the rater selection process.

The following hypothesis related to research question 4 has been tested.

Hypothesis 4: Accuracy of multirater feedback ratings from peers is independent of the favorability of the overall rating.

Since a ‘strongly agree’ response is the only response that equates to a fully committed answer, the responses were coded into two possible categories: ‘strongly agree’ (coded as “positive”) and ‘less than strongly agree’ (coded as “negative”).

Table 38. Cross tabulation for H4: Favorability of Rating

	Accurate		
	Positive	Negative	Total
Positive	22	7	29
Negative	24	63	87
Total	46	70	116

Table 39. Chi-Square Test for H4: Favorability of Rating

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	21.183	1	.000

With a p-value of .000, which is less than .05, the hypothesis is rejected. It can be concluded that the accuracy of feedback ratings depends on the ‘favorability’ of the overall rating, as reported by raters. The more ‘favorable’ or ‘positive’ the rating assigned by the rater, the more likely they were to identify their ratings as “accurate.” This means that some raters attribute the inaccuracy of their ratings to having assigned adverse or unfavorable ratings to the ratee.

Table 40. Chi-Square Test Summary for H1-H4

Variable	Pearson Chi-Square sig. value / p = .05	Significant? Yes or No?
Hypothesis 1a: Gender of Rater	.984	No
Hypothesis 1b: Gender of Ratee	.751	No
Hypothesis 1c: Gender Difference	.648	No
Hypothesis 1d: Age of the Rater	.185	No
Hypothesis 1e: Age Difference	.308	No
Hypothesis 1f: Tenure of the Rater	.403	No
Hypothesis 1g: Confidentiality	.009	Yes
Hypothesis 1h: Comfort Level	.000	Yes
Hypothesis 1i: Opportunity to Observe	.000	Yes
Hypothesis 1j: Training	.139	No
Hypothesis 1k: Purpose	.455	No
Hypothesis 1l: Familiarity of Tasks	.000	Yes
Hypothesis 1m: Nonperformance Factors	.535	No
Hypothesis 2a: Friendship	.003	Yes
Hypothesis 2b: Likeability	.199	No
Hypothesis 2c: Competition	.266	No
Hypothesis 2d: Familiarity with Ratee	.015	Yes
Hypothesis 3: Selection Process	.701	No
Hypothesis 4: Favorability of the Rating	.000	Yes

Table 40 presents all of the statistically significant relationships between specific nonperformance related variables and the reported accuracy of feedback ratings found in this study. Seven of the nineteen hypotheses produced results that were statistically significant. The results of this study concluded that the accuracy of feedback ratings depends on the level of concern for anonymity and confidentiality of the rater, the raters' comfort level with the multirater feedback process, the opportunity raters had to observe the ratee, the familiarity of peer raters to the assigned responsibilities and tasks of the ratee, the friendship established between rater and ratee, on how long the rater has known the ratee, and the "favorability" of the overall rating.

Credibility and Validity of Conclusions

Prior to administering the study, a field test was conducted with seven raters to make any adjustments to the survey. All seven individuals communicated that the directions and the survey questions were clear and easily understood.

Conclusion

Seven of the nineteen hypotheses produced results that were statistically significant and were rejected: *Hypothesis 1g:* Accuracy of multirater feedback ratings from peers is independent of the level of concern for anonymity and confidentiality of the rater, *hypothesis 1h:* Accuracy of multirater feedback ratings from peers is independent of the raters' comfort level with the multirater feedback process, *hypothesis 1i:* Accuracy of multirater feedback ratings from peers is independent of the opportunity raters had to observe the ratee, *hypothesis 1l:* Accuracy of multirater feedback ratings from peers is independent of the familiarity of peer raters to the assigned responsibilities and tasks of the ratee, *hypothesis 2a:* Accuracy of multirater feedback ratings from peers is independent of the friendship between rater and ratee, *hypothesis 2d:* Accuracy of multirater feedback ratings from peers is independent of how long the rater has known the ratee, and *hypothesis 4:* Accuracy of multirater feedback ratings from peers is independent of the favorability of the overall rating.

The results of this study concluded that the accuracy of feedback ratings depends on the level of concern for anonymity and confidentiality of the rater, the raters' comfort level with the multirater feedback process, the opportunity raters had to observe the ratee, the familiarity of peer raters to the assigned responsibilities and tasks of the ratee, the friendship established between rater and ratee, on how long the rater has known the ratee, and the "favorability" of the overall rating. Raters who are less concerned about their anonymity and confidentiality, more comfortable with the multirater feedback process, had greater opportunity to observe the ratee, more familiar with the assigned responsibilities and tasks of the ratee, identified as having a friendship with the ratee, knew the ratee for a greater number of years, and assigned more "favorable" or "positive" ratings, were more likely to identify their ratings to as "accurate." This means that some peer raters acknowledge and attribute rating inaccuracy to factors other than performance.

CHAPTER 5. RESULTS, CONCLUSIONS, AND RECOMMENDATIONS

This chapter explores the results, conclusions, and recommendations resulting from a research study to determine variables that may affect peer raters' responses to a multirater feedback instrument. Numerous variables anticipated to affect peer rater response to an MRF instrument have been examined. Seven of the nineteen hypotheses were rejected. The results of the analysis and research suggest that certain variables, other than performance, are related to the relative accuracy of multirater feedback provided by peer raters. Presented, first, within this chapter, are the research questions and supporting hypotheses followed by a summary of conclusions based on the results of the study. Recommendations for future research will complete the presentation of this chapter.

Research Questions

This research study intended to answer a series of four research questions focused on variables that may influence the accuracy of multirater feedback ratings assigned by peers. The first research question and set of thirteen hypotheses were concerned with the demographics of the peer rater and/or ratee in relation to the accuracy of multirater feedback ratings assigned by peers, including: (a) gender, (b) age, (c) tenure, (d) concern for anonymity / confidentiality, (e) comfort with process, (f) opportunity to observe, (g) training, (h) purpose of assessment, (i) task association/familiarity, and (j) other potential nonperformance factors. The second research question and four corresponding hypotheses examined: (a) friendship, (b) likeability, (c) competition, and (c) how long the rater has known the ratee, in relation to the accuracy of multirater feedback ratings assigned by peers. The third research question and corresponding hypothesis examined the selection process for peer raters in relation to the accuracy of multirater feedback ratings assigned by peers. The fourth, and final, research question and corresponding hypothesis examined the favorability of the overall rating in relation to the accuracy of multirater feedback ratings assigned by peers.

This study addressed four research questions through the development of relevant hypotheses. Statistical techniques were used to either support or not support a total of nineteen hypotheses. The following research questions were examined in this study:

Research Question 1: What is the relationship between the accuracy of multirater feedback ratings from peers versus the nonperformance factors / demographics of the rater and ratee?

To support this question in answering the existence of a relationship between the accuracy of ratings from peers versus nonperformance factors or demographics of the rater and ratee, thirteen hypotheses were developed representing: (a) gender, (b) age, (c) tenure, (d) concern for anonymity / confidentiality, (e) comfort with process, (f) opportunity to observe, (g) training, (h) purpose of assessment, (i) task association / familiarity, and (j) other potential nonperformance factors compared against the constant variable of rating accuracy.

Research Question 2: What is the relationship between the accuracy of multirater feedback ratings from peers versus the personal relationship between the rater and ratee?

To support this question in answering the existence of a relationship between the accuracy of ratings from peers versus the personal relationship between the rater and ratee, four hypotheses were developed representing: (a) friendship, (b) likeability, (c) competition, and (c) how long the rater has known the ratee compared against the constant variable of rating accuracy.

Research Question 3: What is the relationship between the accuracy of multirater feedback ratings from peers versus the selection process for peer raters?

To support this question in answering the existence of a relationship between the accuracy of ratings from peers versus the selection process for peer raters, one hypothesis was developed representing the rater selection process compared against the constant variable of rating accuracy.

Research Question 4: What is the relationship between the accuracy of multirater feedback ratings from peers versus the favorability of the overall rating?

To support this question in answering the existence of a relationship between the accuracy of ratings from peers versus the favorability of the overall score, one hypothesis was developed representing the favorability of the overall rating compared against the constant variable of rating accuracy.

Hypotheses

A total of 19 hypotheses were tested to answer the four research questions examining potential factors and/or variables that may influence the accuracy of multirater feedback ratings from peers. Thirteen hypotheses were developed representing: (a) gender, (b) age, (c) tenure, (d) concern for anonymity / confidentiality, (e) comfort with process, (f) opportunity to observe, (g) training, (h) purpose of assessment, (i) task association / familiarity, and (j) other potential nonperformance factors compared against the constant variable of rating accuracy:

Hypothesis 1a: Accuracy of multirater feedback ratings from peers is independent of the gender of the rater. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings depends on the gender of the rater.

Hypothesis 1b: Accuracy of multirater feedback ratings from peers is independent of the gender of the ratee. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings depends on the gender of the ratee.

Hypothesis 1c: Accuracy of multirater feedback ratings from peers is independent of whether the rater and ratee are the same gender. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings depends on the gender difference between the rater and ratee.

Hypothesis 1d: Accuracy of multirater feedback ratings from peers is independent of the age of the rater. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings depends on the age of the rater.

Hypothesis 1e: Accuracy of multirater feedback ratings from peers is independent of the difference in age between the rater and ratee. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings depends on the difference in age between the rater and ratee.

Hypothesis 1f: Accuracy of multirater feedback ratings from peers is independent of the tenure of the rater with the organization. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings depends on the tenure of the rater with the organization.

Hypothesis 1g: Accuracy of multirater feedback ratings from peers is independent of the level of concern for anonymity and confidentiality of the rater. This hypothesis was rejected. There was sufficient evidence to conclude that the accuracy of feedback ratings depends on the level of concern for anonymity and confidentiality of the rater. Some raters attribute the inaccuracy of their ratings to their concern for anonymity and confidentiality.

Hypothesis 1h: Accuracy of multirater feedback ratings from peers is independent of the raters' comfort level with the multirater feedback process. This hypothesis was rejected. There was sufficient evidence to conclude that the accuracy of feedback ratings depends on the raters' comfort level with the multirater feedback process. Some raters attribute the inaccuracy of their ratings to their lack of comfort with the multirater feedback process.

Hypothesis 1i: Accuracy of multirater feedback ratings from peers is independent of the opportunity raters had to observe the ratee. This hypothesis was rejected. There was sufficient evidence to conclude that the accuracy of feedback ratings depends on the opportunity raters had to observe the ratee. Some raters attribute the inaccuracy of their ratings to their lack of opportunity to observe the ratee.

Hypothesis 1j: Accuracy of multirater feedback ratings from peers is independent of the influence of rater training. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings depends on the influence of rater training.

Hypothesis 1k: Accuracy of multirater feedback ratings from peers is independent of the purpose of peer ratings. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings depends on the purpose of peer ratings.

Hypothesis 1l: Accuracy of multirater feedback ratings from peers is independent of the familiarity of peer raters to the assigned responsibilities and tasks of the ratee. This hypothesis was rejected. There was sufficient evidence to conclude that the accuracy of feedback ratings depends on the familiarity of peer raters to the assigned responsibilities and tasks of the ratee. Some raters attribute the inaccuracy of their ratings to their lack of familiarity of the responsibilities and tasks of the ratee.

Hypothesis 1m: Accuracy of multirater feedback ratings from peers is independent of nonperformance factors. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings was influenced by factors other than performance.

Four hypotheses were developed representing (a) friendship, (b) likeability (c) competition, and (c) how long the rater has known the ratee against the constant variable of rating accuracy:

Hypothesis 2a: Accuracy of multirater feedback ratings from peers is independent of the friendship between rater and ratee. This hypothesis was rejected. There was sufficient evidence to conclude that the accuracy of feedback ratings depends on the friendship established between rater and ratee. Some raters attribute the inaccuracy of their ratings to their lack of friendship with the ratee.

Hypothesis 2b: Accuracy of multirater feedback ratings from peers is independent of how well liked the ratee is by the rater. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings was influenced by how well liked the ratee is by the rater.

Hypothesis 2c: Accuracy of multirater feedback ratings from peers is independent of the degree of competition that exists between the rater and the ratee. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings was influenced by the degree of competition that exists between the rater and the ratee.

Hypothesis 2d: Accuracy of multirater feedback ratings from peers is independent of how long the rater has known the ratee. This hypothesis was rejected. There was sufficient evidence to conclude that the accuracy of feedback ratings depends on how long the rater has known the ratee. Some raters attribute the inaccuracy of their ratings to their unfamiliarity with the ratee.

One hypothesis was developed representing the process for rater selection against the constant variable of rating accuracy:

Hypothesis 3: Accuracy of multirater feedback ratings from peers is independent of the rater selection process. This hypothesis was not rejected. There was insufficient evidence to conclude that the accuracy of feedback ratings depends on the rater selection process.

One hypothesis was developed representing the favorability of the overall rating against the constant variable of rating accuracy:

Hypothesis 4: Accuracy of multirater feedback ratings from peers is independent of the favorability of the overall rating. This hypothesis was rejected. There was sufficient evidence to conclude that the accuracy of feedback ratings depends on the “favorability” of the overall rating, as reported by the raters. Some raters attribute the inaccuracy of their ratings to having assigned adverse or unfavorable ratings to the ratee.

Conclusions

The results of this study will provide researchers, human resource managers, and practitioners with perspective and insight as to the results of multirater feedback systems. This may contribute to more effective interpretation and utilization of MRF assessments. For example, examining how relationships affect peer raters’ use of the rating scale across performance dimensions provides insight into perspectives and biases that may influence ratings.

Specifically, this mixed methods study used an online survey to gather data on how participants responded to a MRF assessment as peer raters. Some researchers (e.g., London & Smither, 2002) have argued that research on MRF has not kept pace with practice and that there are insufficient research models and data available to guide organizations in the use of this type of feedback (Waldman & Atwater, 1998). By studying the factors that influence MRF, human resource managers and practitioners can become more knowledgeable regarding the proper utilization, advantages, and limitations of this type of assessment. The results will build upon this existing research and initiate additional research.

The accuracy of feedback ratings depends on: the level of concern for anonymity and confidentiality of the rater, the raters' comfort level with the multirater feedback process, the opportunity raters had to observe the ratee, the familiarity of peer raters to the assigned responsibilities and tasks of the ratee, the friendship established between rater and ratee, how long the rater has known the ratee, and the "favorability" of the overall rating. This knowledge allows practitioners to properly utilize and interpret the data collected from peer raters in multirater feedback systems.

The research questions explored in this study will contribute to existing research aimed to improve the efficiency of multirater feedback systems. With this knowledge, multirater feedback systems can be designed to minimize or alleviate the identified influences. For example, understanding that peer raters who: are less concerned about their anonymity and confidentiality, more comfortable with the multirater feedback process, had greater opportunity to observe the ratee, more familiar with the assigned responsibilities and tasks of the ratee, identified as having a friendship with the ratee, knew the ratee for a greater number of years, and assigned more "favorable" or "positive" ratings, were more likely to identify their ratings as "accurate", contributes to the interpretation of peer ratings.

Recommendations for Future Research

This study focused on peer rater feedback within multirater feedback systems and the influence of a vast set of nonperformance variables. Future studies delving deeper into one or more of these variables would yield critical information to better understand the influence of nonperformance factors on performance ratings provided by peer raters. The more insight and information that can be gained through research into the variables influencing this feedback, human resource managers and practitioners can become more knowledgeable regarding the proper utilization, advantages, and limitations of this type of assessment.

REFERENCES

Antonioni, D. (1994). The effects of feedback accountability on upward appraisal ratings. *Personnel Psychology, 47*, 349-356.

Antonioni, D., & Park. H. (2001). The effects of personality similarity on peer ratings of contextual work behaviors. *Personnel Psychology, 54*(2), 331-360.

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement: Does it really matter*? Personnel Psychology*, 51, 577-598.

Atwater, L., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self-and follower ratings of leadership. *Personnel Psychology*, *48*, 34-59.

Bernardin, H. J. (1992). An 'Analytic' Framework for Customer-Based Performance Content Development and Appraisal. *Human Resource Management Review,* 81-102.

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

Bloor, M., Frankland, J., Thomas, M., & Robson, K. (2001). Focus groups in social research. Thousand Oaks, CA: Sage.

Blumer, H. (1969). Symbolic Interactionism: Perspective and Method. Berkeley, CA: University of California Press.

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

Bracken, D. W., & Church, A. H. (1997). Advancing the state of art of 360-degree feedback. *Groups and Organization Management, 22*, 149-161.

Bracken, D. W., Timmreck, C. W., & Church, A. H. (2001). *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes*. San Francisco, CA: Jossey-Bass Publishers.

Brutus, S., Fleenor, J., & London, M. (1998). Elements of effective 360 degree feedback. In W. M. Tornow, M. London, and CCL Associates (Eds.), *Maximizing the value of 360-degree feedback* (pp. 11-27). San Francisco, CA: Jossey-Bass Publishers.

Burke, R. J., Weitzel, W., & Weir, T. (1978). Characteristics of effective employee performance review and development interviews: Replication and extension. *Personnel Psychology*, 31, 903-919.

Church, A. H. & Braken, D. W. (1997). Advancing the state of art of 360-degree feedback. *Groups and Organization Management*, *22*, 149-161.

Creswell, M. B. (1963). Effects of confidentiality on performance ratings of professional health personnel. *Personnel Psychology*, 16, 385-393.

DeNisi, A., & Kluger, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

DeNisi, A. S., Randolph, W. A., & Blensoc, A. G. (1983). Potential problems with peer ratings. *Academy of Management Journal*, 26(3), 457-464.

Dominick, P. G., Reilly, R. R., & McGourty, J. W. (1997). The effects of peer feedback on team member behavior. Group & Organization Management, 22, 508-520.

Dorfman, P. W., Stephen, W. G., & Loveland, J. (1986). Performance appraisal behaviors: Supervisor perceptions and subordinate reactions. *Personnel Psychology*, 39, 579-597.

Druskat, V. U., & Wolff, S. B. (1999). Effects of timing and developmental peer appraisals in self-managing work groups. *Journal of Applied Psychology*, 84, 58-74.

Edwards, M., & Ewen, A. (1996). *360 feedback*. AMACOM: New York.

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

Farh, J. L., Cannella, A. A., & Bedeian, A. G. (1991). Peer ratings, the impact of purpose on rating quality and user acceptance. *Group and Organizational Studies*, 16, 367-386.

Fedor, D. B., Bettenhausen, K. L, & Davis, W. (1999). Peer reviews: Employees' dual roles as raters and recipients. *Group & Organization Management, 24*(1), 92-120.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7, 117-140.

Fiske, A. P. (1993). Social Errors in Four Cultures: Evidence about Universal Forms of Social Relations. *Journal of Cross-Cultural Psychology,* 24, 463-494.

Frey, J. H, & Fontana, A. (1991). The group interview in social research. *The Social Science Journal,* 28, 175-187.

Frick, A., Bächtinger, M. T., & Reips, U-D. (1999). *Financial incentives, personal information and drop-out rate in online studies*. In U-D. Reips et al. (Eds.), Current Internet science. Trends, techniques, results.

Gall, M. D., Gall, J. P., & Borg, W. R. (2003). Educational Research: An Introduction (7th ed.). Boston: Allyn & Bacon.

Gay, L. R., Mills, G. E., & Airasian, P. (2006). Educational Research: Competencies for Analysis and Applications. (8th ed.). NJ: Pearson.

Guion, R. M. (1965). Personnel Testing. New York: McGraw-Hill.

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology,* 11, 385-398.

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

Harris, M., & Schaubroeck, J. (1998, March). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings, *Personnel Psychology*, 41(1), 43-62.

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

Hedge, J. W., Borman, W. C., & Birkeland, S. A. (2001). History and development of multisource feedback as a methodology. In D. Bracken, C. Timmreck, & A. Church (Eds.), *Handbook of multisource feedback* (pp. 15-32). San Francisco: Jossey-Bass.

Hillman, L. W., Schwandt, D. R., & Bartz, D. E. (1990). Enhancing staff members' performance through feedback and coaching. *Journal of Management Development, 9*(3), 20-27.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64, 349-371.

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

Kanter, R. M. (1989). When giants learn to dance. New York: Touchstone.

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

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

Koeck, R., & Guthrie, G. M. (1975). Reciprocity in impression formation. *Journal of Social Psychology*, 54, 31-41.

Krueger, R. A. (1988). Focus groups: A practical guide for applied research. Newbury Park, CA: Sage Publications.

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

Lepsinger, R. & Lucia, A. (1997). *The art and science of 360 feedback*. San Francisco, CA: Jossey-Bass Pfeiffer.

Larson, Jr., J. R. (1989). The dynamic interplay between employees' feedback-seeking strategies and supervisors' delivery of performance feedback. *Academy of Management Review*, 14, 408-422.

Latham, G. P. (1989). Job performance and appraisal. In C. L. Cooper & J. Robertson (Eds.), *International review of industrial and organizational psychology*, (117-155). New York: Wiley.

Lewin, A. Y., & Zwany, A. (1976). Peer nominations: A model, literature critique and a paradigm for research. *Personnel Psychology,* 29, 423-447.

Lofland, J., & L. H. Lofland. (1984). Analyzing social settings: A guide to qualitative observation and analysis. (2d ed.) Belmont, CA: Wadsworth.

London, M. (2001). *How people evaluate others in organizations*. Mahwah, NJ: Lawrence Erlbaum.

London, M. (2003). *Job feedback: giving, seeking, and using feedback for performance improvement* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

London, M., & Beatty, R. W. (1993). 360-degree feedback as a competitive advantage. *Human Resource Management, 32*(2-3), 353-372.

London, M., & Smither, J. W. (2002, Spring). Feedback orientation, feedback culture, and the longitudinal performance management process. *Human Resource Management Review, 12*(1), 81-100.

London, M., & Smither, J. W. (1995). 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), 803-840.

Longnecker, C. O., & Goff, S. J. (1990). Why performance appraisals still fail. *Journal of Compensation and Benefits*, 6(3), 36-41.

Love, K. G., (1981). Comparison of peer assessment methods: reliability, validity, friendship bias, and user reaction. *Journal of Applied Psychology*, 48, 211-214.

Matens, J. (1991). Conducting effective performance reviews. *Modern Casting,* 89(6), 54-46.

McEvoy, G. M. & Buller, P. F. (1987). User acceptance of peer appraisal in an industrial setting. *Personnel Psychology*, *40*(4), 785-797.

Meyer, H. H., Kay, E., & French, J. R. (1965). Split roles in performance appraisal. *Harvard Business Review*, 43, 123-129.

Mirbella, J. (2006). Hypothesis testing with SPSS: A non-statistician's guide & tutorial. Author. Available from http://www.drjimmirabella.com/ebook/

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

Mohrman, S. A., Cohen, S. G., & Mohrman, A. M. Jr., (1995). *Designing team based organizations: New forms for knowledge work.* San Francisco: Jossey-Bass.

Morgan, D. L. (1997). Focus groups as qualitative research. London: Sage.

Mount, M. K., & Scullen, S. E. (2001). Multisource feedback ratings: What do they really measure? In M. London (Ed.), *How people evaluate others in organizations* (pp. 155-176). Mahwah, NJ: Erlbaum.

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., & Mohler, C. J. (2001). Reliability, validity, and meaningfulness of multisource ratings. In D. W. Bracken, C. W. Timmreck, & A. H. Church (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (pp. 130-148). San Francisco: Jossey-Bass.

Nadler, D. A. (1979). The effects of feedback on task group behavior: A review of the experimental research. *Organizational Behavior and Human Performance*, 23, 309-338.

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

Nowack, K. M. (1993, January). 360-degree feedback: The whole story. *Training and Development, 47*(1), 69-72.

Oberg, W. (1999). Make performance appraisal relevant. Retrieved on April 4, 2009 http://www.zigonperf.com/resources/pmnews/pas-relevant.html

Passini, F. T., & Norman, W. T. (1966). A universal conception of personality structure? *Journal of Personality and Social Psychology*, 4, 44-49.

Peterson, D. E., & Hillkirk, J. (1991). A better idea. Boston, MA: Houghton Mifflin.

Pitkow, J. E. & M. M. Recker (1994). Results from the first world-wide web user survey. Retrieved on April 5, 2009 http://www.cc.gatech.edu/gvu/user_surveys/survey-04-1995/

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

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

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

Roberts, G. E. (1998). Perspectives on enduring and emerging issues in performance appraisal. *Public Personnel Management*, 27(3), 301-313.

Romano, C. (1994). Conquering the fear of feedback. *HR Focus*, 71, 9-19.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsh, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, 37, 407-422.

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

Scott, W. D., Clothier, R. C., & Spriegal, W. R. (1941). Personnel Management. New York, NY: McGraw-Hill.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85, 956–970.

Shaver, W., Jr. (1995). *How to build and use a 360-degree feedback system.* Alexandria, VA: American Society for Training and Development.

Spriegal, W. R. (1962). Company practices in appraisal of managerial performance. *Personnel*, 39, 77.

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

Timmreck, C.W., & Bracken, D.W. (1997). Multisource feedback: A study of its use in decision making. *Employment Relations Today*, 24, 21–27.

Tornow, W. W. (1993, Summer-Fall). Perceptions or reality: Is multi-perspective measurement a means or an end? *Human Resources Management*, *32*(2), 221-229.

Tsui, A. S., & Barry, B. (1986). Interpersonal affect and rating errors. *Academy of Management*, 29, 586-589.

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

Van Velsor, E. V., Leslie, J. B., & Fleenor, J. (1997). *Choosing 360: A guide to evaluating multirater feedback instruments for management development*. Greensboro, NC: Center for Creative Leadership.

Varma, A., Denisi, A. S., & Peters, L. H. (1996). Interpersonal affect and performance appraisal: A field study. *Personnel Psychology*, *49*(2), 341.

Waldman, D.A., & Atwater, L.E. (1998). *The power of 360 degree feedback: How to leverage performance evaluations for top productivity*. Houston, TX: Gulf Professional Publishing Company.

Wexley, K. N., & Klimoski, R. (1984). Performance appraisal: An update. In K. M. Rowland & G. R. Ferris (Eds.), Research in personnel and human resources management, 2, 35-79. Greenwich, CT: JAI Press.

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

Whisler, T. L., & Harper, S. F. (1962). Performance appraisal: Research and practice. New York, NY: Holt, Rinehart, and Winston.

Wiese, D. S., & M. R. Buckley. (1998). The evolution of the performance appraisal process. *Journal of Management History*. 4(3), 233-249.

Zalesny, M. D. (1990). Rater confidence and social influence in performance appraisal. *Journal of Applied Psychology*, 75, 274-289.

APPENDIX. RESEARCH QUESTIONS AND SURVEY QUESTIONS

Research Question(s)	Hypotheses	Survey Question(s)
Research Question 1. What is the relationship between the accuracy of multirater feedback ratings from peers versus the nonperformance factors / demographics of the rater and ratee?	Hypothesis 1a: Accuracy of multirater feedback ratings from peers is independent of the gender of the rater.	2
	Hypothesis 1b: Accuracy of multirater feedback ratings from peers is independent of the gender of the ratee.	3
	Hypothesis 1c: Accuracy of multirater feedback ratings from peers is independent of whether the rater and ratee are the same gender.	2 and 3
	Hypothesis 1d: Accuracy of multirater feedback ratings from peers is independent of the age of the rater.	4
	Hypothesis 1e: Accuracy of multirater feedback ratings from peers is independent of the difference in age between the rater and ratee.	5
	Hypothesis 1f: Accuracy of multirater feedback ratings from peers is independent of the tenure of the rater with the organization.	9
	Hypothesis 1g: Accuracy of multirater feedback ratings from peers is independent of the level of concern for anonymity and confidentiality of the rater.	10
	Hypothesis 1h: Accuracy of multirater feedback ratings from peers is independent of the raters' comfort level with the multirater feedback process.	11
	Hypothesis 1i: Accuracy of multirater feedback ratings from peers is independent of the opportunity raters had to observe the ratee.	12
	Hypothesis 1j: Accuracy of multirater feedback ratings from peers is independent of the influence of rater training.	16
	Hypothesis 1k: Accuracy of multirater feedback ratings from peers is independent of the purpose of peer ratings.	6
	Hypothesis 1l: Accuracy of multirater feedback ratings from peers is independent of the familiarity of peer raters to the assigned responsibilities and tasks of the ratee.	17

	Hypothesis 1m: Accuracy of multirater feedback ratings from peers is independent of nonperformance factors.	20
Research Question 2. What is the relationship between the accuracy of multirater feedback ratings from peers versus the personal relationship between the rater and ratee?	Hypothesis 2a: Accuracy of multirater feedback ratings from peers is independent of the friendship between rater and ratee.	13
	Hypothesis 2b: Accuracy of multirater feedback ratings from peers is independent of how well liked the ratee is by the rater.	14
	Hypothesis 2c: Accuracy of multirater feedback ratings from peers is independent of the degree of competition that exists between the rater and the ratee.	15
	Hypothesis 2d: Accuracy of multirater feedback ratings from peers is independent of how long the rater knows the ratee.	8
Research Question 3. What is the relationship between the accuracy of multirater feedback ratings from peers versus the selection process for peer raters?	Hypothesis 3: Accuracy of multirater feedback ratings from peers is independent of the rater selection process.	7
Research Question 4. What is the relationship between the accuracy of multirater feedback ratings from peers versus the favorability of the overall rating?	Hypothesis 4: Accuracy of multirater feedback ratings from peers is independent of the favorability of the overall rating.	18