

**MULTISOURCE PERFORMANCE FEEDBACK:  
COMPARISON WITH NORMATIVE RATING PATTERN**

By

Margo Parker Hudson

SUZANNE KAVLI, Ph.D., Faculty Chair

DAVID NOTGRASS, Ph.D., Committee Member

SANDRA REID, Ph.D., Committee Member

Dean, Gary Cook School of Leadership

A Dissertation Presented in Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

Dallas Baptist University

Summer 2011

UMI Number: 3503563

All rights reserved

INFORMATION TO ALL USERS

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

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



UMI 3503563

Copyright 2012 by ProQuest LLC.

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



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

© Margo Parker Hudson, 2011

We hereby recommend that the submitted Dissertation

Title: *Multisource Performance Feedback: Comparison with Normative Rating Pattern*

By: *Margo Parker Hudson*

Be accepted in partial fulfillment of the requirements for the Degree of:

**Doctor of Philosophy**

---

Ph.D. Program Director

Date

**Dissertation Committee**

---

SUZANNE KAVLI, Ph.D., Committee Chair

Date

---

DAVID NOTGRASS, Ph.D., Committee Member

Date

---

SANDRA REID, Ph.D., Committee Member

Date

**Dean of Graduate School**

---

Gary Cook School of Leadership

Date

## Abstract

Leadership performance assessments, commonly called 360 degree feedback, multisource feedback, or multirater feedback, are used by organizations in an effort to guide leaders to develop their skills. However, as the process has become popular with numerous organizations, questions about rater accuracy have risen. While research has examined influences on rater accuracy in assessment results, no studies were found that compared archival multisource feedback data with a normative pattern. Consequently, this study explored the following question: “What is the nature of the differences between performance ratings from different rating sources?” The present study is significant to the field of research regarding multisource feedback because it compiled source-specific feedback and demonstrated that when the results were displayed by source, it is possible to interpret the meanings of the competency ratings by source. This quantitative study compared the correlation matrix of actual data with a normative pattern derived from a questionnaire administered to a different team of executives to determine if there are differences between the source groups, as well as the bases for differentiating the rates. The study revealed that the correlation matrices of all the source types and the normative structure differed significantly. The absolute magnitudes of the differences between the correlation matrices, both among the source types and between the normative and source type groups, were measured by computing the Euclidean distances of the correlations. The research found that the raters’ rating policy differences made no difference in the ordering of the scores, meaning that differences between the raters were almost completely due to the differences in the way they actually rated people. The results

presented statistically significant evidence that the structure of the various sources differ and that specific differences exist between the normative factor structure and the source type structures for self, peers, and manager and for the grouping of direct reports and customer/clients. Further, from a practical perspective, researchers and professionals may be able to improve the interpretation of rater feedback in multisource feedback evaluations. The result would be meaningful feedback with more accuracy for the ratees, facilitators, and coaches to use in development planning.

Keywords: multisource feedback, normative rating pattern, performance assessment, performance evaluation, performance appraisal, source differences, multirater feedback, 360 feedback, accuracy

## **Dedication**

Dan Hudson is my business partner and my life partner, and, since I began this journey four years ago, he has been my support partner in the pursuit of the doctorate degree from Dallas Baptist University. Throughout our marriage of 33 years, Dan has looked at challenges with the attitude of, “How can we make this work?” Not only have my studies required enormous amounts of time, but travel from our home near Springfield, Missouri, to Dallas has added to the time I have had to be away from Dan and home. Nonetheless, through serious illnesses and family issues these past four years, Dan’s encouragement of my doctoral work has never faltered. When I was wrestling with statistics courses, I was especially happy that I had the foresight to marry a math major. To you, dearest Danny, I dedicate this volume. Together, we have made it work.

## Acknowledgments

With gratitude, I acknowledge the following individuals for their support and contributions throughout this process:

My husband, Dan, and my children, grandchildren, and dearest friends whose constant encouragement and support strengthened my resolve to complete my goals.

My friend and editor, Louise Gilbreath, whose dedication to this project made possibility a reality.

Dr. Sue Kavli, chair of my committee, who agreed to undertake this journey by aligning with the vision of what this study could be and providing guidance, support, expertise, and kindness to me throughout the process.

Dr. David Notgrass, who through his suggestions and his expertise in Leadership Studies strengthened the project.

Dr. Sandra Reid, who through her continuous support, guidance, and encouragement helped me understand the significance of this study for Leadership Development and Human Resource Management.

Dr. Jeffrey Kane, who shared his expert advice and encouraged me to have a worthy study by generously giving his time throughout the methodology process.

The faculty members and staff of Dallas Baptist University's Gary Cook Leadership Studies, led by Dr. Jeremy Dutschke, who regularly provided encouragement and support.

Cohort III of Dallas Baptist University's Ph.D. in Leadership Studies program, who provided me with support, mentoring, and friendship during our learning time together.



Stan Yee, my longtime friend and former industry colleague, for his generous gifts of time and expertise in converting numbers to meaningful graphics.

The senior staff of EADS—Jussi Itävuori, Hervé Borensztejn (formerly of EADS), Nancy Gröesch (formerly of EADS), David Fink, David R. Oliver, Jr., Sean O’Keefe, and Ralph Crosby, Jr.—who without hesitation agreed to the use of EADS’ multisource feedback archival data that formed one leg of this study.

Michael Moody, Randy Hutcherson, senior leaders, and Amanda Luzader, support staff, of Force Protection, Inc., who provided access and support to me for gathering data from 32 Force Protection leaders whose responses on a leadership competencies questionnaire formed the normative data, the second leg for this study.

My brother, Forrest Mosten, and my sister-in-law, Dr. Jody Mosten, for their love and continuous support.

Dr. Joel Persky, who has since passed away, encouraged me to strive for excellence and to continue my education. I miss him.

Inez and Charles Wright, who provided me warm hospitality and made their home, my home during my academic journey.

Michael Schwarz, my longtime colleague and dear friend, who reviewed the document with keen eyes and offered helpful comments.

Claudia Wilson, my cousin, who gave her valuable time to read the document with “fresh eyes.”

## Table of Contents

Acknowledgments	iv
CHAPTER 1. INTRODUCTION	1
Introduction to the Problem	1
Background of the Study	3
Statement of the Problem	3
Purpose of the Study	6
Rationale	7
Research Questions	9
Significance of the Study	9
Definition of Terms	12
Assumptions and Limitations	14
Organization of the Remainder of the Study	15
CHAPTER 2. LITERATURE REVIEW	17
CHAPTER 3. METHODOLOGY	51
Introduction	51
Research Design	52
Target Population and Sample	54
Instrument	55
Validity and Reliability	57
Data Collection	66
Data Analysis Plan	67
Ethical Considerations	76



## List of Tables and Figures

Table 1. 2x2 Table of Competency Ratings	68
Table 2. Normative Correlation Matrix	82
Table 3. Actual Data Matrix	83
Table 4. Numbers and Descriptive Statistics of Rates per Competency Dimension by Rater Type	85
Table 5. Results of Comparisons of Competency Dimension Correlation Matrices between Source Types	88
Table 6. Results of Comparisons of the Normative and Source Type Competency Dimension Correlation Matrices	90
Table 7. Euclidean Distances Between the Lower Triangular Portions of the Source Type Correlation Matrices	92
Table 8. Euclidean Distances Between the Lower Triangular Portions of the Normative and Source Type Correlation Matrices	93
Table 9. t-test Results for Mean Differences Between Each Source Type's Actual Overall Mean Scores and Those Formed With the Weights of Each Other Source Type	98
Table 10. Comparisons between Pairs of Most Similar Rotated Factors between Source Types	99
Table 11. Comparisons between Pairs of Most Similar Rotated Factors for the Normative Factor Solution and that for Each Source Type	106
Figure 1. Examples of 360 Degree Feedback Instruments and Processes	4
Figure 2. Leadership Competencies	7

Figure 3. Discussion of Summary Results	121
Figure 4. Source Group: Self, Manager and Peer	123
Figure 5. Source Group: Direct Reports & Customer/Client	124
Figure 6. Source Group: Normative	127

## CHAPTER 1: INTRODUCTION

### Introduction to the Problem

Leadership performance assessment processes have been widely researched and published in numerous books, articles, and dissertations. Organizations desiring to assess the performance of their senior leaders and high potential employees for professional development often use a feedback process to gather data from multiple sources in the leader's sphere of influence. Nowack (1993) gave five reasons for the increased use of 360 degree systems for feedback:

- a need for cost-effective alternatives to assessment centers
- the increasing availability of assessment software capable of summarizing data from multiple sources into customized feedback reports
- the need for continuous measurement in continuous improvement efforts
- the need for job-related feedback for employees affected by career plateauing
- the need to maximize employees' potentials in the face of technological changes, competitive challenges, and increased workforce diversity. (p. 35).

Jones and Bearley (1996) held that 360 degree feedback is necessary as organizations change strategies and organizational structure. Typically, professionals within the training, organizational development, talent management, or leadership development departments, or independent management consultants and development coaches administer the feedback instrument.

Although there are numerous versions of feedback instruments with various names, they are similar in format, content, and process. Some of the common names are multisource feedback, multirater feedback, or 360 degree feedback, used

interchangeably in this document. Bracken (1994) found the 360 degree or multirater feedback to be dependable and consistent. It is easy to use, relevant, and understandable and can cause change for the better both for individuals and organizations. Heisler (1996) said that the 360 degree feedback process is useful for measuring employee behavior and individuality. Generally, the individual being rated, called the ratee or the participant, selects people, called raters, to assess the ratee's leadership competencies by providing feedback on a written instrument. The raters typically include self, subordinate, peer, manager, and customer/client.

Bracken (1994) said that a good multirater feedback process would include “process design and planning, instrument design and development, feedback processing and reporting, action planning and follow up” (p. 46). The ratee receives the compiled feedback from a facilitator or career coach who helps interpret the results and assists the ratee in creating a career development plan. Dalton (1996) proposed that feedback assessments be used only for employee development, not for pay delivery decisions. The use of multisource feedback for compensation decisions could produce invalid or negative results (Dalton, 1996).

Figure 1 lists various brands of multisource feedback instruments. Briefly described here are three instruments commonly used by U.S. businesses: LPI (Leadership Practices Inventory), Benchmarks, and PROFILOR®. The first, LPI, is based on Kouzes and Posner's Five Practices of Exemplary Leadership® model. The five practices are (a) Model the Way, (b) Inspire a Shared Vision, (c) Challenge the Process, (d) Enable Others to Act, and (e) Encourage the Heart (Leadership Practices Inventory Web site, 2011; Van Velsor & Leslie, 1991). Benchmarks is the product

of the Center for Creative Leadership (Center for Creative Leadership Web site, 2011; Van Velsor & Leslie, 1991). It measures 16 competencies that the center calls “critical for success,” as well as five possible career spoilers. The instrument asks raters how they perceive the ratee’s strengths and development needs, how the ratee ranks with managers in other organizations, and what skills and perceptions are critical to success. The results are the basis for a development plan that links the ratee’s needs to specific development experiences. Personnel Decisions International sells the PROFILOR®, a 360 degree feedback process that provides an evaluation of the ratee’s strengths and development needs (PDI Web site, 2011).

### **Background of the Study**

The assumption in using a multisource feedback instrument is that each rater (self, subordinate, peer, manager, and customer/client) has a unique perspective on the rated individual's leadership competencies. The rationale for these multisource processes is that feedback that combines the unique assessments of several individuals give the rated individual a more nearly complete picture of his or her performance. With the guidance of a facilitator/coach, the ratee then can initiate a professional development plan.

### **Statement of the Problem**

When using multisource performance feedback, the question arises: Are the ratings accurate, or are there other considerations besides actual performance that affect the ratings? Matens (1999) stated that only 20 percent of performance evaluations are viewed as meeting the requirement of being part of an effective organizational system for performance management. Additionally, multisource ratings may be no more accurate



than a typical manager-employee performance appraisal (Fletcher, Baldry, & Cunningham-Snell, 1998, p. 479). Wherry and Bartlett (1982) looked for causes other than performance that affect ratings and examined methods that might eliminate or minimize biases. They also proposed factors that affect the accuracy of ratings, including job performance, biases, and measurement error. A study by Scullen, Mount, and Goff (2000) categorized the effects of five factors that supposedly influenced ratings: (a) the ratee's overall performance, (b) the ratee's performance on a specific competency, (c) the rater's individual rating tendencies, (d) the rater's organizational awareness, and (e) random error (p. 956). Kane (2004) believed that the results of conventional appraisal methods are "typically so distorted by bias and so difficult to compare across jobs and even raters, that they fail to adequately serve their intended purposes" (p. 2). Accuracy of the feedback instrument is key to the ratee's acceptance of the results and its value as a development tool. By comparing actual archival data from one international company's multisource feedback process with a normative pattern derived from a questionnaire completed by a group of senior leaders of a second company, this researcher examined the relationship between two correlation matrices and explored the possible interpretations for the patterns of ratings.

<b>Figure 1. Examples of 360 Degree Feedback Instruments and Processes</b>		
<b>Product</b>	<b>Provider</b>	<b>Description</b>
LPI	Pfeiffer	Provides for self-assessment, feedback from supervisors, co-workers, direct reports, and others with direct knowledge of the ratee in a leadership role
PROFILOR®	Personnel Decisions International	Feedback tools and processes that offer individual development suggestions. Determines strengths and development needs

<b>Figure 1. Examples of 360 Degree Feedback Instruments and Processes (continued)</b>		
<b>Product</b>	<b>Provider</b>	<b>Description</b>
360 Feedback and Assessments 20/20 Insight GOLD	CRG Associates Resources Unlimited Workplace Results, LLC	Feedback from a number of people about performance of an individual, team, or organization
360 Degree Inventory Assessments, Profiles	Executive Development Associates A Performance and Learning Strategies Group	
Checkpoint 360	Training Solutions, Inc.	
CheckPoint 360 Feedback Assessment	Integrated Management Resources	Web-based assessment that assures confidentiality, measures 70 items, can have up to 10 respondents
Circumnavigator	HR Coaching Tools	30 categories and survey questions; allows users to write their own competencies and survey questions and/or rating scales and rater roles
DISC Personal Profile System	Goeins-Williams Associates, Inc.	Assesses behavioral styles of organizational members and provides feedback
Discovering Diversity Profile	Corexcel	Identifies reactions to cultural diversity and develops understanding. It helps simplify issues, increase understanding of differences, identify possible conflict and transform information into acceptance
Individual Performance in Three Dimensions Assessment	Learning LAB Associates	Performance feedback from the individual, coworkers & management

**Figure 1. Examples of 360 Degree Feedback Instruments and Processes (continued)**

<b>Product</b>	<b>Provider</b>	<b>Description</b>
Relating and Communicating Style Assessment	Roberts & Roberts Associates	Contrasts self-perception with that of those who know the ratee well
Performance Management and Employee Score Cards	The Rainmaker Group	Tracks and manages employee performance
Personal & Behavior Assessments	Chatsworth Consulting Group	Personal and behavior assessments and feedback tools that help ratee be more successful in career and life
MLQ 360 Assessment Report (Multifactor Leadership Questionnaire)	MLQ International, a Division of MLQ P/L, Melbourne	Nine leadership styles (45 items) assessed at the individual level

### **Purpose of the Study**

The study examined the relationship between the normative pattern and archival pattern of ratings and interpreted the findings. The researcher compared the memory-based competency correlation matrix of actual data with the normative pattern to determine the basis of the differences between source groups and what may be the bases for differentiating the rates. The study used a conceptually derived normative correlation matrix of competencies made up of the same competencies that EADS used in the multisource feedback process as shown in Figure 2.

## **Figure 2. Leadership Competencies**

### **A. Lead Courageously**

Build a vision and give clear strategic direction; make tough decisions and take responsibility for them; inspire and influence others; act with integrity

### **B. Drive Excellence & Innovation**

Reliably deliver results by executing on time, cost, and quality; foster innovation, change, and continuous improvement

### **C. Develop Yourself & Others**

Challenge people to grow; actively develop people; learn continuously

### **D. Communicate Honestly & Effectively**

Ensure direct and clear dialogue; proactively address conflicts and problems

### **E. Deliver in Our Global Environment**

Demonstrate and apply international business knowledge; get things done through networks in and outside EADS; be a team player across cultures and organizations

### **F. Generate Customer Value**

Focus the organization on customer needs and expectations; manage customer relationships

## **Rationale**

With limited success, previous studies have examined rater differences and the nature of source differences in multisource feedback. Murphy and DeShon (2000) stated that there are at least four reasons that account for rater differences in performance appraisals other than the reason of random measurement error: (a) “systematic differences in what is observed,” (b) “systematic differences in access to information other than observations of performance,” (c) “systematic differences in expertise in interpreting what is observed,” and (d) “systematic differences in evaluating what is observed” (p. 882).

Borman (1974), Murphy and Cleveland (1995), and Tornow (1993) noted that raters observe different behaviors of ratees from their unique opportunities to view such behaviors. Lance, Woehr, and Fisicaro (1991) demonstrated that source effects associated with performance ratings from raters who are at different levels indicate true performance variance (Lance & Woehr, 1989) and sometimes the raters are privy to observe differing amounts of behavior (Wherry & Bartlett, 1982). Campbell and Lee (1988) speculated that different conceptual frameworks may influence how raters rate ratees. Raters at locations different from the ratee's workplace may have uneven opportunities to observe behaviors and know about the ratee's successes and failures (Bernardin & Beatty, 1984; Murphy & Cleveland, 1995; Murphy & DeShon, 2000). Hoffman (2006) indicated that some source differences may simply represent variance that cannot be attributed to the ratee's performance. This study used archival data and a normative pattern to compare source differences and interpret the meaning of source factors to add to the research data about source differences.

Hoffman (2006) examined the meaning of multisource feedback source effects and their correlation with external constructs. His research indicated that multisource feedback source effects do reflect performance-relevant variance and that the performance-relevant variance is source specific (p. 13). Waldschmidt (2006) concurred with London and Smither (1995), who said that "it is clear that multisource feedback is an area in which practice is well ahead of theory and empirical research" (p. 807). Waldman, Atwater, and Antonioni (1998) pointed to the lack of knowledge on "how or even whether 360 feedback really works" (p 89). Based on his own findings, Hoffman (2006) encouraged future researchers to further examine the causes of multisource

feedback source effects so that the continued use of multisource feedback would be meaningful.

### **Research Questions**

This study posed the following research question, “What is the nature of the differences between performance ratings from different rating sources?”

1. Are the differences between the sources’ correlation matrices for the actual data statistically significant?
2. Are the differences between the sources’ correlation matrices for the actual data source types vs. the correlations for the normative group statistically significant?
3. What are the relative magnitudes of the differences between the correlation matrices from different sources?
4. How much difference do implicit performance theories make in the total rating scores that ratees receive?
5. How do the structures of the matrices for various sources differ?
6. How similar is the rotated factor structure (of the maximum number of non-error principal components) to the similarly rotated factor structure of the normative group?

### **Significance of the Study**

The study is significant in that it examined whether multisource feedback results from one international company are true reflections of what the raters think and how their opinions match with a normative pattern from a second international company. Previous research has provided rationale for source differences with no definitive conclusions.

Consequently, the nature and interpretation of the meaning of source differences needs additional inquiry. The findings of this study will also help to clarify whether there are differences between various sources. If there are differences between the sources, are they significant, and what interpretations may account for the nature of differences? The findings have the potential to clarify source specific feedback in multisource feedback assessments, resulting in more meaningful feedback with greater accuracy for ratees and coaches to use for development planning. The results of a comparison of archival data with a normative pattern may encourage leaders to be more cognizant and attuned to the design, interpretation, and implementation processes for multisource feedback within their organizations. Limited research has been published regarding the nature of the differences between sources. The research posits that the pattern of correlation on items of memory-based multisource feedback instruments may reveal more about each source type's cognitive framework than about "what" correlates with "what" across individual differences in performance (Shweder & D'Andrade, 1980).

Used interchangeably in this paper, three terms are employed for the research methods that use "judges" to answer questions on a conceptually based questionnaire. The terms are *conceptual association matrix*, *evaluative salience matrix*, and *normative pattern matrix*. The researcher used a normative or conceptual pattern of six leadership competencies (Figure 2) derived from a group of executives' responses to compare with several years of another company's archived 360 degree feedback data. The archival actual data came from the international company, European Aeronautic Defence and Space Company, subsequently referred to in this document as EADS. EADS uses a 360 degree feedback instrument to examine the leadership competencies of its senior

executives. The EADS multisource feedback instrument is a customized and proprietary adaptation of PROFILOR®, a widely used multisource feedback instrument. The University of Minnesota in its Global Leadership Solutions continuing education course uses PROFILOR® to assess students' leadership competencies and to guide the creation of a professional development plan for them. "The PROFILOR® is a proven tool for gathering...input. Your personal and confidential results, an integral part of the development experience, will help you improve, grow, and develop as a leader" (SMLP Web site, 2011).

EADS, organized just over a decade ago from a group of European aerospace and defense companies, calls itself "a global leader in aerospace, defense and related services" with operations and markets on every continent (EADS Web site, 2011). In 2010, EADS employed 120,000 people worldwide. Its products include jet and turboprop military and commercial aircraft, helicopters, space systems, defense and security systems, and integrated security solutions, such as border security, maritime security, and crisis response.

The EADS data show a pattern in the ratings that can be compared with the normative model. The normative pattern was derived from a questionnaire used to obtain ratings of a hypothetical grouping of competencies completed by the senior leaders of a second international defense company; Force Protection, Inc. Force Protection, Inc. provides survivability solutions to support the armed forces of the United States and its allies. According to the company's Web site:

Force Protection designs, manufactures, tests, delivers, and supports its blast- and ballistic-protected products to increase the survivability of the users of the



products. Its specialty vehicles are designed to protect their occupants from landmines, hostile fire, and improvised explosive devices. The company is a key provider of the U.S. military's mine resistant ambush protected vehicle program and has sold and delivered more than 3,000 vehicles under the program. (Force Protection Web site, 2011)

The U.K. Ministry of Defense is also a customer. Force Protection is based in Ladson, South Carolina, with offices in Michigan, South Carolina, Virginia, and the United Kingdom. The company also has a repair and support operation in Kuwait and provides similar support in Afghanistan and Iraq (Force Protection Web site, 2011)

### **Definition of Terms**

The terms used in this research study are defined here.

*Bias:* raters' inability to judge rates on the competencies being assessed without other factors influencing the ratings (Landy & Farr, 1980).

*Construct validation (external):* the covariance between sources and competencies and constructs assessed by an external measurement system.

*Developmental purposes or career development:* compiled feedback used for planning personal or professional improvement of the ratee.

*Feedback:* information a rater sends or a ratee receives about the ratee's leadership competencies.

*Implicit theories:* conceptual schema about phenomena that exist in the mind of the rater. Raters' implicit theories may affect performance ratings.

*Interrater agreement:* the level of agreement within and across rating sources.

*Internal approaches to construct validation:* the examination of the relationship of raters from different levels of an organization using a common instrument to rate participants.

*Latent structure of multisource feedback:* a variable that is determined through questions posed by the researcher. Structural modeling is a method of evaluating performance latent structure.

*Measurement equivalence:* an internal approach to construct validity; signifies that the instrument functions identically across raters (Cheung, 1999; Vandenberg & Lance, 2000).

*Multirater, multisource, and 360 degree feedback:* A process for obtaining performance ratings from multiple individuals on more than one level of an organization (Borman, 1997; Harris & Schaubroeck, 1988; Latham, 1999).

*Peer:* a coworker with similar experience or organizational level.

*Performance appraisal, performance evaluation, or performance review:* a formal employee evaluation process that assesses a ratee for possible administrative use or developmental planning or both.

*Ratee:* an individual being evaluated or assessed.

*Rater:* an individual completing and returning an assessment instrument.

*Rating policy:* refers to judgment policy; or the weights a judge attaches to multiple judgment criteria, or the competencies, in arriving at an overall judgment. In the present study, the term means the relative weights assigned to the competencies by a source type.

*Reliability:* “The extent to which a set of measurements is free from random-error variance” (Guion, 1965, p. 30).

*Validity:* the determination of how well an instrument measures what it purports to measure (Guion, 1980).

### **Assumptions and Limitations**

**Assumptions.** Assumptions are points of view that the researcher regards as fact. Assumptions are the basis of the overall research. They can influence the type of research methods used and, consequently, can affect the study’s validity. In this study, there is an assumption that, in answering a multisource feedback survey, each rater (self, subordinate, peer, manager, and customer/client) has a unique perspective on the rated individual's leadership competencies. The rationale for these multisource processes is that feedback that combines the unique assessments of several individuals gives the rated individual a more nearly complete picture of his or her performance. Another assumption is that the Force Protection executives who made up the conceptual group thoroughly understood the questions and answered thoughtfully and honestly. The study also assumed that participants were objective when they answered the survey questions. Since the data for the actual ratings came from another organization, the questionnaire for Force Protection executives was about hypothetical constructs wherein the participants, ratees, and their organizations remained anonymous and unidentifiable. The study also assumed that the number of Force Protection participants was adequate and appropriate.

**Limitations.** The limitations of a study involve the areas that could not be investigated nor explained sufficiently by use of the instrument, analysis, or interpretation of the data gathered. One limitation of this study was that, post distribution of the

questionnaire to Force Protection participants, the researcher was unable to alter the questions or clarify items for the raters. Further, although the EADS multisource feedback instrument is based on PROFILOR®, it is nonetheless a customized instrument and, perhaps, unique. Instructions to EADS participants are given “online” with no additional training for the raters. Additionally, translation of competencies from English to other languages may have complicated the ratings since some raters’ native language is not English. Further, the participants picked the raters with the potential of influencing rating accuracy because of an association between rater and ratee. Yet another limitation was that there may have been a concern or suspicion regarding a questionnaire related to the subject of performance appraisal or performance evaluation. Additionally, questionnaires such as the one distributed to Force Protection executives cannot probe deeply into the psyches and experiences of the raters (Gall, Gall, & Borg, 2003). Notwithstanding these concerns, the findings of this research will be beneficial to companies that use multisource feedback.

### **Organization of the Remainder of the Study**

Chapter two presents a review of the literature regarding multisource feedback, source differences, and studies that use a normative pattern for comparison with archival data. Chapter three submits the methodology for the research project. Using statistical methods described later, the researcher compared the actual competency correlation matrix for the full data set and for each rating source—self, subordinate, peer, manager, and customer/client—to a conceptually derived normative matrix. Chapter four includes the analysis and results of the data collected. Chapter five includes the summary of the findings and interpretations of the results, including limitations, implications, and

recommendations. The paper concludes with the references and appendices containing correspondence, copies of the instrument and the questionnaire, and other significant support documentation used as part of the study.

## CHAPTER 2: LITERATURE REVIEW

### Overview

Organizations carry out leadership assessments in an attempt to assess how their leaders are viewed by their managers and colleagues (Bartram, 2004, p. 238) so that the ratee can improve his or her leadership competencies. As a way to improve the performance assessment of their employees, a number of Fortune 500 companies have adopted a multisource feedback process (London & Smither, 1995). The multisource feedback assessment process provides performance feedback to an individual by collecting input from multiple individuals within the ratee's sphere of influence. The basis for collecting multiple surveys is that the combined feedback from those rater sources will give the rated individual a more nearly complete picture of his or her performance and/or developmental gaps, leading to improved performance. Each rated individual receives a report that shows self ratings compared to peer, boss, subordinate, and customer/client ratings. Those ratings form the basis of coaching and planning sessions.

As a basis for any research project, Boote and Beile (2005) favored a literature review with several important objectives: "It sets the broad context of the study, clearly defines what is and is not within the scope of the research, and justifies those decisions. It also situates existing literature in a broader scholarly and historical context" (p. 4). They pointed out that an examination of the literature allows the researcher to discern what has been accomplished and what remains to be done in a given field of study. "[A] good literature review," they wrote, "is the basis of both theoretical and methodological

sophistication, thereby improving the quality and usefulness of subsequent research” (p. 4).

In the spirit of Boote and Beile’s 2005 delineation of a literature review, this researcher examined the literature on multisource or 360 degree performance feedback and comparisons with a normative pattern. The review also included the history of 360 degree or multisource feedback assessments in organizations, the rationale for multisource instead of single-source performance assessments, the meaning and use of feedback, and rater accuracy. The researcher also explored the facets of rater accuracy as found in the literature; that is, interrater agreement and disagreement, internal construct validity, measurement equivalence, latent structure, external construct validity, and implicit theories and biases affecting rater judgments.

The voluminous literature on multisource feedback speaks to its importance and wide use in organizations seeking to continuously improve the capability of their senior leadership. As a leadership development practice, multisource feedback requires work and time, and the process, which is dependent on the variables of human input and interpretation, has flaws. Yet, researchers found for the most part that 360 degree feedback is superior to other forms of performance rating (Bozeman, 1997; Kane, 2004).

### **What Is Multisource or 360 Degree Performance Feedback?**

Dalessio (1998) described performance feedback as “evaluations gathered about a target participant from two or more rater sources, including self, supervisor, peers, direct reports, internal customer/clients, external customer/clients, and vendors or suppliers” (p. 278). The literature reveals the historical use of multisource feedback, its strengths and weaknesses, and the rationale for its use as a method to assess leadership

performance. Although multisource feedback is a popular tool in management development, especially in large organizations (London & Smither, 1995), opinions and the empirical research on its efficacy as a development tool are uneven.

A multirater feedback instrument usually is based on a scaled survey distributed to multiple individuals chosen by the person to be rated on the skills and abilities considered essential to performing a job effectively (Figure 2). The method of gathering and assessing the data can be formal or informal (McCauley & Van Velsor, 2004, p. 6). Responders rate each competency on a scale, for example, of 1 to 5, with 5 meaning “all the time” and 1 meaning “never,” to indicate how often they believe that the ratee behaves in that particular manner (Fleenor & Prince, 1997, p. 71). The process usually includes a self-appraisal as well as ratings by subordinate, peer, manager, and customer/client (Mohrman, Resnick-West, & Lawler III, 1990).

The completed instruments are returned to a facilitator/coach, who often is a consultant hired to coach the ratees. The facilitator/coach compiles the multiple responses into a report and in a private session helps the ratee interpret results. The report typically comments on strengths, weaknesses, and areas for improvement, and the coach identifies and guides reflection on disconnects between the ratee’s self-appraisal and others’ assessments. Then the ratee and coach establish a development plan that includes goal-setting based on the organization’s strategic development plan. Ideally, there is a midyear review of progress and an adjustment of goals as appropriate. At the end of 12 months, an evaluation of the process is conducted.



## **History of Multisource or 360 Degree Performance Feedback**

Conceptually, the multisource performance appraisal process originated from several theoretical arenas. The theory of social cognitive theory (Bandura, 1986; Stajkovic & Luthans, 1998) states that individuals have the insight to perform meaningful self-reflections on their own capabilities—leading to self-awareness. The multisource feedback process provides an avenue for structured input from rater sources in addition to self, thus increasing one’s self-awareness of elements that need to be included in a development plan (Lomakina, 2008, p. 34). Further, the heart of assessment is “the conversation that takes place” first within the individual as he or she receives information from the world, then with others as they share their worlds, then between the coach and the participant as the person learns the world of the “experts,” and lastly between the individual and his or her work (McCauley & Van Velsor, 2004, p. 28).

Some progressive companies have used such assessments for decades. Waldman, Atwater, and Antonioni (1998) traced the multisource feedback process to the 1950 and 1960 human relations movement. In the 1970s, Clark Wilson, an industrial psychologist, introduced the multisource feedback concept to management training (Wilson, 2003). Wilson’s research and development of early survey instruments transformed the way that businesses evaluate their managers. The use of multisource performance feedback in creating development plans increased significantly in the decade of the 1990s. In 1990, Mohrman, Resnick-West, and Lawler had only two chapters in their book discussing “multisource appraisals.” They focused on who appraised performance and what legal considerations applied when feedback was used for pay and promotion. In 1993, McGarvey and Smith wrote that upward feedback was growing in popularity and

advocated a structure to maximize benefits “with the least amount of pain” (p. 28). Also in 1993, London and Beatty found that multisource feedback could result in performance improvements and a competitive advantage for the organization. By 1997, Brutus, Fleenor, and London declared that multisource performance feedback had become “one of the most popular methods ... in organizations...to improve the effectiveness of their managers” (p. 6). Brutus et al. (1997) listed several examples of Fortune 500 companies, including UPS, AT&T, Amoco, General Mills, and Procter & Gamble, that were either implementing multisource feedback or had investigated its use for their organization.. Other researchers who reported on the increasing use of multisource feedback included London and Smither (1995) and Yammarino and Atwater (1997).

### **Feedback**

The purpose of rater feedback on multisource instruments is to assess a participant, give feedback, and help him or her interpret the information in order to form a development plan (Chory & Westerman, 2009; Geddes, 1993). Feedback is generally confidential, and raters remain anonymous. Feedback usually is presented individually (London & Smither, 1995), often by a trained coach/facilitator hired for that purpose, and the compiled results are used to develop a plan for professional development and growth. Yukl (2006) called feedback one of the primary influences on leader traits. The immediate task in a feedback-intensive experience is to analyze the information from the multiple raters on the ratee’s style and behavior. Yet in any feedback experience, the challenge is not just confronting the individual with data about personality, behavior, or performance from many rater sources; the additional challenge is accepting the multiplicity of views that often differ from one’s self-view and one from the other.

Frequent sharing of informal feedback, especially in manager-subordinate coaching relationships, can lead to advantageous employee performance (Hellqvist, 2008).

Research indicates that raters who receive negative feedback often use that feedback to create a development plan (Bailey & Austin, 2006; Brutus et al., 1997; Maurer, Barbeite, & Mitchell, 2002).

If the feedback is free of bias, the ratee can compare the self-assessment ratings to others' ratings, consider the extent to which others' ratings agree with each other, and revise his or her self-image accordingly (London & Smither, 1995). If intentional and unintentional biases exist, the coach and the ratee should be aware of the potential for the biases to color the responses and, therefore, should interpret the results in the context of the biases. Wherry and Bartlett (1982) indicated that identifying and interpreting biases allows for methods to be developed to improve the assessment process.

A primary advantage of multisource feedback is its value to individuals who might not otherwise receive such information. The rated individuals can see how those who know them well perceive them. Illuminating a ratee's blind spots is essential for development (Wimer & Nowack, 1998, p. 74). Wimer and Nowack (1998) also concluded that when raters are given the chance to rate participants without attribution, they are sometimes able to rid themselves of years of unexpressed feelings. They hope that their leaders will change. The authors also said that a group's morale and success often improve dramatically after the members have articulated their opinions and know that someone is taking them seriously.

Researchers have determined that participants who invest themselves in the process of multisource feedback tend to learn more about themselves and have a more

meaningful evaluation process (Rosti & Shipper, 1998, p. 77). Also, the ratee may gather valuable insights from people with different perspectives (Yammarino & Atwater, 1993). The ratee views feedback as more meaningful if it comes from both supervisors and direct reports rather than from direct reports alone. The ratee is also more likely to accept developmental feedback if the rater sources include both coworkers and direct managers (Bernardin, Dahmus, & Redmon, 1993). London and Smither (1995) agreed: The ratee may show more interest and desire to create his/her development plan if the feedback is viewed as coming from influential raters. The process of multisource feedback may provide organizations an assessment tool that is viewed as fair by the participants. The process has the potential “to provide greater reliability, enhanced fairness, and increased ratee acceptance” (Thiry, 2009, p. 21). Multisource feedback also provides an organization with an opportunity for participants to accept and want to use the appraisal system (Wexley & Klimoski, 1984). The process has gained support because it eliminates some negative aspects of single manager only assessments.

### **Multisource Assessments Versus Single-Source Assessment**

Before multisource feedback processes became accepted, performance appraisals for the most part consisted only of the manager of the employee being reviewed. The direct supervisor or manager of the employee was responsible for the performance appraisal (Murphy, Cleveland, & Mohler, 2001) because the direct supervisor or manager was viewed as the most informed about the competency level of his/her employee. Managers, however, were reluctant to perform such evaluations, and even recent research revealed that feeling has not changed: Managers often dread the process and view conducting appraisals as challenging. Consequently, supervising managers postpone

performance reviews and modify their feedback, especially when they have negative information to share (Tjosvold & Halco, 1992, p. 629). Benedict and Levine (1988), as well as Gruenfeld and Weissenberg (1966), also found that managers do not enjoy assessing employees' performance. Similarly, if managers are not knowledgeable about current instances of performance that allows them to evaluate an employee's behavior accurately, their ratings may be inaccurate and biased (Longnecker & Goff, 1990). London (2003) concurred: "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). Greenberg (1986a, 1986b) held similar views. Multisource feedback processes have greater accuracy and objectivity than single-source processes (Thiry, 2009, p. 2). Researchers found that the multiple ratings of raters with differing perspectives and roles led to increased organizational acceptance as raters with unique perspectives evaluated the ratee (Borman, 1997; Harris & Scharbroeck, 1988; Latham 1999). Reliability and validity of multisource feedback may be viewed as more valid by the ratee than feedback received from a direct supervisor or manager (Mount, Judge, Scullen, Sytsma, & Hezlett, 1998).

Bartram (2004) saw multisource feedback as an intervention to assist selected individuals in improving leadership skills and to measure the potential and performance of selected employees. Wimer and Nowack (1998) wrote that multisource feedback, when done well, may lead to increased team and organizational effectiveness and individual behavioral change. Favorable reactions to the feedback are necessary for positive behavior change and should result in the recipient's seeking additional feedback

from raters and setting developmental goals to ensure leadership development (Facteau, Facteau, Schoel, Russell, & Poteet, 1991).

In addition to benefits for the individual, an organization can realize positive changes from using multisource feedback (Wimer & Nowack, 1998). The process can indicate to the employee what competencies are valued in the culture, thus enabling the ratee to develop in that direction. With that knowledge, the ratee can plan for professional development to coincide with the organization's plans. The use of multisource feedback systems also demonstrates that the organization wants and values employee feedback. Dialogue between the participant and raters as a result of multisource feedback may improve both relationships and the working environment, resulting in a better functioning organization (Wimer & Nowack, 1998).

Shaver (1995) also suggested that multisource feedback helps people:

...uncover expectations, strengths, and weaknesses that are news to them; it broadens the perspective on evaluating an individual by using multiple data rater 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)

It is appropriate here to discuss the terms *error* and *bias* as they appear in this paper. Kruglanski and Ajzen (1983) drew a distinction between "bias" and "error." Bias may not become error, they wrote, if the latter term means "a departure from some accepted criterion of [external] validity" (p. 18). Thiry (2009) defined bias as the "systematic tendency for ratings to be influenced by anything other than the behavior

being measured” (p. 14). Funder (1987) wrote that an error is “any deviation of an individual observation from a standard, such as a sample or population mean, or the output from a predictive model, for example, a regression” (p. 78). For this study, error refers to any variance that is not related to a ratee’s “true performance.” In other words, error can be random error or systematic bias that is unrelated to a target's true performance (Hoffman, 2006, p. 45).

### **Rater Accuracy**

As part of the literature review, the researcher explored rater accuracy, one of the concerns about the use of multisource feedback (Landy & Farr, 1980). Despite considerable research comparing multisource feedback assessments to single-source supervisory ratings, there remains limited understanding of what may influence accurate ratings by raters who are not on the leadership team or who haven’t been trained to assess performance (Thiry, 2009). Matens (1999) reported that the percentage of performance appraisals that accurately reflect work force performance is quite low (20%).

Kane (2004) held:

Even the appraisal methods considered most rigorously developed, such as behaviorally anchored rating scales, are seriously flawed by documented deficiencies. Rating errors occur in two forms: intentional and unintentional. Intentional rating errors are a rater’s deliberate distortions to make ratings conform to some preconceived notion about the ratee’s performance. (p. 3)

If the ratee and the coach could trust that the ratings are factual, a leadership development plan for the subject would be more valuable. However, the reviewed studies leave open the question of rater differences and what makes ratings from different

rater sources different. This researcher found no studies that compared archival multisource feedback data with a normative pattern. Thiry (2009) wrote that within the process of multisource performance feedback bias exists because, in rating participants, raters apply their observations, judgments, and insights. In the evaluation of performance, rating accuracy is affected by biases and judgmental error.

The literature review considers facets of rater accuracy, including interrater agreement and disagreement, internal construct validity, measurement equivalence, latent structure, external construct validity, and biases affecting rater judgments. The literature review also looks at the following source-related differences in ratings: (a) interrater agreement and disagreement, (b) measurement equivalence, (c) structural modeling, and (d) intentional and unintentional biases. The practical implication of understanding rater differences and the nature of source differences is that an executive can ill afford the time and effort to work on developing the wrong skills or run the risk of ignoring a major deficiency.

### **Interrater Agreement and Disagreement**

In examining the relationship between a normative pattern and an archival pattern of findings, the study explored interrater agreement and disagreement, including the pattern, if any, of differences between rater sources. Bradley (2004) found that “correlations between self- and other ratings are modest at best” (p. 23). On the other hand, Harris and Schaubroeck (1988) found that peers and supervisors tend to agree more than other source pairs. Additionally, there seems to be limited understanding as to why the other source pairs have low agreement, according to Tsui and Ohlott (1988). They



also wrote that disagreement challenges assumptions about source differences as well as differences between sources and ratees.

Antonioni and Park (2001) specifically investigated “the relationship between the raters’ interpersonal affect toward leaders and the leniency of ratings from 360 degree feedback” (p. 480). For example, they asked whether liking a ratee affects the ratings. They concluded that a highly visible relationship between the ratee and manager may be a factor in the ratings because raters are able to witness the interaction between the parties. A positive relationship with the manager may also influence the raters’ insight into a ratee’s leader’s behavior. In other words, raters who have strong impressions about the ratee may be especially attentive to the behaviors that reflect their judgments. Subordinates and peers, rather than the supervisor, are more influenced by the ratees’ pattern of interactions with others (Antonioni & Park, 2001). Similarly, Pulakos, Schmitt, and Chan (1996) suggested that the interactions and behaviors that the rater is able to observe cause the disagreement between rater levels. They also suggested that “raters from different organizational levels may consider different factors in formulating their ratings of leaders” (p. 117).

Bozeman (1997) wrote that, according to the common concepts of reliability and validity, when interrater agreement is low, it may point to a lack of reliability and validity in the process (p. 313). However, some scholars have contended that ratings from different rater groups may be valid even if they do not have high levels of agreement. Further, since raters within various groups observe an individual's behaviors under different circumstances, one could expect disagreement among rater groups. Additionally, he said that ratings from different groups of raters likely correlate with the

respective roles of the ratee and the rater. That is, various rater groups are inclined to evaluate the ratee's competencies that relate most to the raters themselves. For example, the ratee's direct management may evaluate a ratee differently from the way the ratee's direct reports rate him or her (Bozeman, 1997, p. 314). Bozeman also said that reliability and validity in the sense of multisource appraisals are basically irrelevant, and interrater agreement among different rater groups is not essential to the validity of the ratings (p. 314).

Bozeman (1997) further suggested that although it does not seem appropriate to require interrater agreement across rater groups, it seems logical for purposes of reliability and validity of the ratings to require agreement within rater groups (p. 314). He said that if the evaluations of raters within a given rater group exhibit little systematic agreement or concurrence pertaining to the behavior of the rated individual, then the reliability, and, therefore, the validity, of the ratings should be seriously questioned. Interrater agreement within rater groups would enable assessments of rating validity to be conducted. He also suggested that different formats and criteria be developed for different rater groups. "It does not seem logical to expect supervisors, subordinates, and peers to respond similarly to identical generic questions about an individual's performance" (Bozeman, 1997, p. 315).

### **Internal Approaches to Construct Validation**

Hoffman (2006) conducted a statistical analysis between an assessment center (an external source) versus (internal) multisource feedback data to study why there might be differences in ratings among rater sources. Hoffman's research compared actual ratings from a multisource feedback assessment with the actual ratings from an assessment

center study. He found a wealth of multisource feedback research on the amount of agreement within and across rater sources to establish the psychometric properties of ratings from multiple rater sources, including Conway and Huffcutt (1997); Harris and Shaubroeck (1988); and Viswesvaran, Ones, and Schmidt (1996). Hoffman said that the research consistently revealed that the evaluators who shared the same organizational level rated more similarly than those from different organizational levels. However, he found research that showed systematic differences in ratings related to the organizational level of the rater (Hoffman, 2006, p. 12). Hoffman (2006) concluded that the research offers “initial evidence for the importance of considering rater organizational level when interpreting performance rating data” (p. 12).

Borman (1974, 1997), Hoffman (2006), and Woehr et al. (2005) reported that, despite general agreement that both rater sources and the competencies they rate explain the significant variation in multisource feedback ratings, the meaning of the respective effects is open to interpretation. That is, internal approaches to construct validation fail to offer conclusive evidence as to the construct validity of dimension factors or the meaning of source factors (Hoffman, 2006, p. 3). Hoffman (2006) further pointed out that, although research provides useful information as to the quality of multiple ratings, internal approaches to construct validation leave important issues unresolved. Some writers suggested that this body of research still needs to explore interpretations of the meaning of multisource feedback source effects (Farr, 2006). Viswesvaran, Schmidt, and Ones (2002) relied solely on internal approaches to construct validation to provide evidence that source factors mean nothing.

On the other hand, some researchers, for example, Borman (1997), Farr (2006), and Murphy and DeShon (2000), argued that the results of studies incorporating internal approaches should not be relied upon in determining the validity of multisource feedback instruments. In other words, although the internal approaches regularly demonstrate the presence of source within raters and between raters and dimension effects, that is, how it relates to the competency, they do not provide evidence regarding what the variance represents, or, the validity of these effects. Consequently, those authors argued that research relying solely on the internal approaches results in erroneous conclusions with respect to the reliability and validity of multisource feedback (Hoffman, 2006, p. 5). Research indicates, however, that multisource feedback instruments for the most part are valid.

Alimo-Metcalfe (1998) wrote that remarkably consistent themes have emerged from research into interrater agreement and disagreement on the use of multisource performance feedback. In general, she said that ratees self-score leadership skills and behaviors more favorably than do other raters. Additionally, she noted that direct reports are more satisfied with their direct supervisor or manager and have higher job satisfaction when their managers' views match their self perceptions. In a corollary, she pointed out that direct reports have less job satisfaction and are less satisfied with their direct supervisor or manager when the managers' views disagree with their own.

Yammarino and Atwater (1997) listed four categories of self-other appraisal agreement: (a) Over-estimators. The over-estimators tend to rate themselves higher than others rate them. Self-ratings tend to be higher because the ratee is unaware of how others view him or her, and raters tend to be reluctant to provide negative feedback

(Yammarino & Atwater, 1997, p. 40). (b) In Agreement/Good. In the *in-agreement/good* category, self- and other ratings tend to be high and to match. Self-raters call themselves “ideal employees,” “good managers,” and “effective leaders” (Yammarino & Atwater, 1997, p. 40). (c) In Agreement/Poor. In measuring *in-agreement/poor* category, self-ratings and others’ ratings tend to be less favorable and dissimilar. In other words, the ratee has self-awareness and recognizes areas that need development. In this category, the ratee is confronted with information that requires an action plan for improvement (Yammarino & Atwater, 1997, p. 40). (d) Under-estimators. The under-estimators’ self-ratings tend to be lower than others’ ratings. Either the ratee is not giving himself credit for accomplishments or is reluctant to rate himself high. When a ratee receives favorable feedback from other raters, that information tends to encourage the ratee to think about the information received and continue to seek ways to change and develop. However, the research indicates that the ratees’ willingness to use feedback for improvement vary, with some ratees accepting and changing and others ignoring the feedback and the opportunity for improvement (Yammarino & Atwater, 1997).

The literature explores the significance of the differences in performance ratings from self, subordinate, peer, manager, and customer/client. Previous studies indicated that there are often differences in the evaluation across raters, as well. Previous studies concentrated on identifying the source of differences or on finding agreement among raters and then examined agreement among the sources (Bradley, 2004, p. 1). On the other hand, Borman (1974) wrote, “Raters at different organizational levels probably observe significantly different facets of a ratee's job performance in most organizations.

If so, their ratings ought to reflect these differences” (p. 105). With data from many people, he said, the rater can reflect on the similar views and the views with significant differences. The researchers exploring the question of differences or agreement among raters included Bradley, 2004; Harris and Schaubroeck, 1988; Maurer, Raju, & Collins, 1998; Mount et al., 1998; Shapira and Zevulun, 1989; Tsui and Ohlott, 1988; and others.

### **Measurement Equivalence**

Measurement equivalence is an internal step toward construct validity. It signifies that the instrument functions identically across sources (Cheung, 1999; Vandenberg & Lance, 2000). If significant comparisons among employees are to be made over time, Greguras (2005) wrote, it is necessary to examine the equivalence of the instruments that evaluate job-related competencies. He found limited research examining measurement equivalence of rating and the different rating sources within organizations (p. 393). Austin and Villanova (1992) noted the consistent lack of researchers’ interest in psychometric factors that influence the ratings. Waldschmidt (2006) found that measurement equivalence research involving multisource feedback ratings produced mixed results with respect to measurement equivalence among rater sources. Rater sources might differ in their conceptualization of the dimensionality of job performance, he wrote, and thus measurement inequivalence might be found for certain performance dimensions and not for others. Although ratings from different rater sources tended to differ, results from both confirmatory factor analysis (CFA) and item response theory (IRT) analysis methods agreed that measurement equivalence was established for both sources and performance competencies.

Greguras (2005) said that instead of research focusing on measurement equivalence, it has focused on the instrument, the performance review process, or factors influencing the accuracy of the ratings. Maurer et al. (1998) found that confirming that measurement equivalence exists is an important step to accomplish before examining ratings between sources. Cheung (1999) reported on two types of equivalence: conceptual and psychometric. Greguras, (2005) wrote that conceptual equivalence indicates that if raters view the questions asked on the instrument in the same way, then their scores will reflect that similar understanding. Conceptual equivalence should be confirmed prior to an analysis of the differences between source groups (Cheung & Rensvold, 1999; Reise, Widaman, & Pugh, 1993). Previous studies have explored how to determine conceptual equivalence for multisource feedback instruments (Facteau & Craig, 2001). Additionally, when raters reply to the questions in the instrument in a similar manner, their scores reflect that similarity by showing equivalent levels of intercorrelation among factors, means, range, reliability, and variance (Greguras, 2005). Therefore, in order to examine differences, conceptual equivalence is needed, and in order to explore rater differences, understanding of how psychometric equivalence functions is required.

Using CFA analysis, Facteau and Craig (2001) concluded that differences in ratings between sources across competencies are equivalent. Hoffman (2006) found that other researchers incorporated a structural modeling approach to evaluating the psychometric properties of multisource feedback scales. Others taking such an approach included Cheung (1999); Conway (1996); Coovert, Craiger, and Teachout (1997); Diefendorff, Silverman, and Greguras (2005); Facteau and Craig (2001); Lance and

Bennett (1997); Lance, Teachout, and Donnelly (1992); Lawler (1967); Maurer et al. (1998); Woehr, Sheehan, and Bennett (2005); and Zedeck and Baker (1972).

In a 1999 study, Cheung pointed out that measurement equivalence researchers had typically specified separate models of ratings from different rater sources and had examined whether rating source moderates the pattern or magnitude of loadings on latent performance factors. In essence, Hoffman (2006) said that this type of examination might show what competencies are contained in a distinct performance factor and which of those competencies rated by different sources may have equivalent loading on latent factors. Researchers have used this type of study to determine the extent to which ratings on a given multisource feedback instrument are comparable across rater sources.

Together, measurement equivalence studies aim to show the degree to which different rater groups picture the dimensions of job performance in a similar way. The examined research, except for a study by Lance and Bennett (1997), constantly showed that performance ratings are equivalent across rater sources (Diefendorff, Silverman, & Greguras, 2005; Fecteau & Craig, 2001; Maurer et al., 1998; Scullen, Mount, & Judge, 2003). Despite the findings of traditional multisource feedback measurement equivalence research, the usual methods of assessing equivalence leave unanswered questions as to the construct validity of multisource ratings (Hoffman, 2006, p. 16).

Another method of examining multisource feedback ratings is to specify models that include both dimension and source effects (Hoffman, 2006). This subject is further explored later in this study. Woehr et al. (2005) assessed the equivalence of scores across rater sources by modeling both dimension and source factors derived from a multitrait-multimethod matrix (MTMM). The results of their study were consistent with prior



equivalence research, suggesting that the primary performance dimension being rated was equivalent across rater sources; whereas, the impact of rater sources differed across rater sources. Hoffman (2006) believed that the approach of Woehr et al. (2005) was a superior method of examining the measurement equivalence of multisource feedback because it specified a theoretically significant model. However, Woehr et al. (2005) drew their sample from military mechanics and engineers, not managers. Hoffman replicated the previous finding, examining source and dimension factors in an industry setting.

### **Latent Structure of Multisource Feedback**

In order to confirm the construct validity of competencies, or to interpret differences between sources, construct validity analysis needs to extend beyond just an internal examination (Hoffman, 2006, p. 3). A latent variable is determined by the researcher through questioning. Structural modeling provides a method of evaluating the latent structure of performance. Hoffman (2006) surveyed a number of rater sources reporting on the latent structure of multisource feedback. The study by Woehr et al. (2005) was the only attempt he found that explored the equivalence of multisource feedback ratings using a model that includes both source and dimension latent factors. Conversely, he noted that a wealth of research has explored the construct validity of performance ratings provided by multiple organizational rater sources from the MTMM perspective. Advances in statistical techniques have resulted in alternate methods of analyzing MTMM data. Of the modern approaches to examining MTMM data, confirmatory factor analysis has received the most widespread use (Goffin & Jackson, 1992). Confirmatory factor analytic methodologies typically specify both trait and method latent factors and also examine the relative factor loadings of dimension and

method effects (Conway, 1996). This approach determines the relative proportion of variance that competencies and methods explain in multisource feedback ratings.

In the context of multisource feedback ratings, the dimension or competency being rated serves as the trait, and the source that provides the rating serves as the method. In essence, multisource feedback ratings represent a special case of the MTMM methodology that can be referred to as a multitrait multisource matrix (MTMS). These approaches begin by generating a matrix containing the correlation (or covariance) among each dimension measured by each rating source. For example, the three dimensions evaluated in the Hoffman (2006) study (conceptual/administrative, interpersonal, and leadership skills) rated by three rater sources (peers, supervisors, and subordinates) resulted in a matrix containing nine lower diagonal correlations, one for each source's rating on each dimension. An early study by Lawler sought to examine the effect of both dimensions and sources multisource performance ratings using a MTMS methodology.

Since Lawler's initial evaluation in 1967 of multisource ratings using the MTMS methodology, significant research has explored the effect of source and dimension factors on ratings from raters who are at different management levels within organizations. Although Hoffman's (2006) research incorporated a variety of methodologies—for example, average HTMM and MTHM correlations and CFA—the results regularly indicated that raters who represented different levels in the organization had ratings that contained both source and dimension effects. Finally, uniqueness (idiosyncratic error) represents the most variance in performance ratings (29%). The findings of Conway's review are in keeping with other research that examined the comparative impact of

dimensions, rater sources, and uniqueness that is found in multisource performance ratings (e.g., Woehr et al., 2005). In summary, research that explored the structure of multisource feedback ratings tended to show that the variance in the ratings from participants across the organization were due to the presence of source and dimension effects.

### **External Approach: Normative Pattern**

Hoffman (2006) conducted a study using data from an assessment center (an external source) versus (internal) multisource feedback data to determine why there might be differences in ratings among rater sources. He found sources, including Borman (1974, 1997), Lance and Woehr (1989), and Lance, Woehr, and Fisicaro (1991), who noted that the usefulness of multisource feedback tools is predicated on different rater sources offering different perspectives of a target's performance. If each source's ratings agreed perfectly, the information from multiple rater sources would be redundant. Similar to the Hoffman study, this study compared a normative pattern derived from a hypothetical questionnaire, an external source, with actual archival data from a multisource feedback assessment, an internal source, to determine why there might be differences in ratings among rater sources.

Those in other fields using the conceptual association or normative pattern to compare with actual data included Bales (1970); Dahlstrom and Welsh (1960); Sears, Macoby, and Levin (1957); and Shweder (1975). Newcomb (1931) used the conceptual association matrix to obtain opinions of "What conceptually goes with what?" in 26 behaviors based on pre-existing rater beliefs. Newcomb (1931) asked 10 University of Chicago students to make paired comparison judgments of conceptual similarity on a 7-

point scale. In this example, 1 is very dissimilar conceptually, and 7 is very similar conceptually. Shweder (1975) examined the four experiments doing re-analysis, matching data with a conceptual schema of “What goes with what?” The experiments included Newcomb’s 1931 test of the validity of a rating technique; Bales’ 1970 personality and interpersonal behavior study; Sears, Macoby, and Levin’s 1957 *Pattern of Child Rearing*; and Dahlstrom and Welsh’s 1960 conceptual similarity handbook.

Shweder (1975) re-examined the four Newcomb (1931) experiments using the following methods: With the 1931 data, he examined the impact of the pre-existing concepts held by the student judges on how similarly they rated behaviors (Shweder, 1975, p. 462). Based on a 1970 experiment by Bales, Shweder placed Bales’ 26 items on cards and presented one each to 50 undergraduate students at Harvard University. The purpose was to obtain judgments of “what conceptually goes with what” based on pre-existing rater beliefs. The students sorted the items into piles according to written instructions. An association measure calculation was determined between each pair of items from the task and these data formed the conceptual association matrix (p. 468).

Sears, Macoby, and Levin’s (1957) 44 scales were converted into declarative assertions and presented to five University of Chicago graduate students. For example, “strictness about bedtime” became “Mother is strict about bedtime.” Using a method similar to Dahlstrom and Welsh’s 1960 experiment, Shweder (1975, p. 474) asked the students to make paired comparison judgments of conceptual similarity on a 7-point scale to obtain ratings of “What conceptually goes with what?”

The study examined both source differences in the hypothetical and actual ratings and examined how much difference implicit performance theories make in the scores that

ratees receive. Rating accuracy is assured when raters are aware of biases affecting a performance evaluation (Thiry, 2009, p. 3). Heslin and VandeWalle (2008) wrote:

Organizational effectiveness requires that personnel be managed, developed, and rewarded based on their actual performance rather than on managers' flawed perceptions of an employee's performance. In reality, however, managers' perceptions and decisions about employees are often systematically biased.  
(p. 219)

### **Intentional and Unintentional Biases**

There are many examples of intentional biases that cause rater error. One is the age of the ratee, wherein younger employees receive higher ratings than older employees who are doing the same job (Ferris, Yates, Gilmore & Rowland, 1985). Another is raters' knowledge of prior performance of the ratee (Huber, Neale, & Northcraft, 1987). Steiner and Rain (1989) also listed in order the observance of good and poor performance. Raters' expectations of the ratee, including supervisors' expectations, also can cause rater error (Hogan, 1987). Expectations of future performance and reluctance to give negative feedback also can play a part in rater error (Larson, 1989). A reliance on "scripts" to give negative performance feedback also can color a rater's decisions about the subject (Dugan, 1989). Greenberg (1986a, 1986b) also found that the perception of organizational fairness can play a role in rater accuracies.

"Unintentional rating errors are inaccuracies in ratings that result from unconscious biases" (Kane, 2004, p. 3). These include stereotyping, halo effect, and systematic distortion, discussed in the following sections. Other examples of

unintentional biases include memory decay (Kozlowski & Kirsch, 1987) and rater intelligence (Smither & Reilly, 1987). Bias may result from a conceptual recollection of the latest behavior incidences that the rater experienced with the ratee; remembering that behavior may influence how the rater scores the ratee (Murphy, Gannett, Herr, & Chen, 1986).

### **Stereotyping**

Stereotyping is a common form of unconscious bias. Stereotypes are a result of assumptions that people make about other people's beliefs and behaviors based on an often erroneous image about what those people may be like. Breslin (1991) indicated that bias based on factors such as race, religion, gender, national origin, and employment status may exist. One study of stereotypes compared the characteristics of Americans and Asians. Americans were seen to be outgoing and arrogant, while Asians were expected to be reserved in nature and shrewd in their dealings. Shweder and D'Andrade (1980) wrote that stereotypes may be deceptive correlations because individuals remember unique characteristics of people in distinctive groups, for example, minorities, more clearly than they can recall behaviors by non-minority individuals. Raters can therefore remember behavior examples of individuals belonging to a distinctive group even when there is no correlation.

### **Halo Effect**

Halo effect is a tendency for positive characteristics to be associated with other positive characteristics. It is also a form of illusory correlation. Research by industrial and organizational psychologists indicates that halo effect accounts for significant rater measurement error in performance evaluations (Mount et al., 1998, p. 560). Mount et al.

(1998) also noted that halo effect on multisource feedback instruments occurs when an overall evaluation of the ratee influences the rater's ratings.

Cooper (1983) observed that halo effect increases as the richness of the description of the quality decreases. He said that groups of rating questions that are “unclear, imprecise, non-specific, ill-defined, and overlapping” continue to be used despite advice to instrument developers to avoid them (p. 490). He also noted that the more alike the qualities are, the less the halo effect. Raters tend to form one impression of different, but similar, categories and rate each question the same. Cooper said that this general impression is created by taking separate behavioral incidents and thinking of them as an accurate description of the ratee’s overall performance. The likely result, he said, is “substantial halo, indicated by high intercategory correlations or low intercategory variance” (Cooper, 1983, p. 491). In an earlier study, Cooper (1981) pointed out that those who work with rating data are cognizant of the halo effect factor. Cooper (1983) reviewed nine methods, for example, rating irrelevant categories, used to lessen illusory halo, concluding that the most effective method being used was to increase the sample of the ratees' actual behavior examples. He also concluded, however, that most rating processes probably result in halo effect in the scores.

### **Systematic Distortion Theory**

Another type of unconscious bias is systematic distortion. Having researched systematic distortion in psychological testing, Shweder and D’Andrade (1980) defined it in this way: “Under difficult memory conditions judges on personality inventories, rating forms, and questionnaire interviews infer what ‘must’ have happened from their general beliefs about what the world is like and/or [they] find it easier to retrieve conceptually

related memory items” (p. 37). The authors further said that one’s general beliefs about what the world is like in the area of personality (one’s implicit personality theories) have a tendency to be inaccurate regarding how behaviors go together. They confuse “what is like what” with “what goes with what.” Stated another way, the systematic distortion hypothesis suggests that judges on memory-based personality procedures are prone to a cognitive illusion in which “propositions about language” may be confused with “propositions about the world” (D’Andrade, 1965, p. 215), and similarity in meaning is mistaken for the likelihood of co-occurrence (Chapman, 1967; Chapman & Chapman, 1967, 1969; Shweder, 1977). Shweder and D’Andrade (1980) also said that the significance of their findings is that there is a pattern to the errors that raters make. Systematic bias may result because some questionnaires require the rater to summarize many categories of judgments over time. The accuracy of the ratings is masked by systematic bias; so, how the rater actually judges performance is uncertain. In other words, because of systematic bias, it is difficult to distinguish individual differences in behavior and rely on rater accuracy (Shweder & D’Andrade, 1980, p. 54).

Shweder and D’Andrade (1980) suggested that the systematic distortion hypothesis can be tested by comparing the similarity between interbehavior patterns of association that come from judgments with similar meanings. They further said that a memory-based hypothesis supposes that correlations of behaviors in memory-based assessments are inaccurate reports about the interbehavior correlations found in actual behavior. Instead, they said, memory-based ratings reveal the level where the labels for the behaviors are similar in meaning.



Shweder and D'Andrade (1980) further said that patterns of correlation among items on memory-based personality instruments reveal more about diverse forms of implicit person theory in the minds of raters than about “what” actually compares with “what” across ratee differences in behavior. They concluded that such forms of conceptual affiliation could be most easily discovered by simply asking a handful of informants, “What is like what?” (p. 38). Examination of the correlational structure of psychiatric ratings using categories from the Brief Psychiatric Rating Scale revealed that this rating structure could be reproduced from judgments about “similarity in meaning.” Thirty minutes of videotaped interaction among members of a family revealed that memory-based rating structures paralleled pre-existing similarity-of-meaning structures but did not accurately reflect the correlational structure of actual behavior (Shweder & D'Andrade, 1980, p. 54).

Camerer (1988) built on Shweder and D'Andrade's (1980) studies in clinical psychology by applying the systematic distortion theory to organizational traits. Camerer (1988) said that people incorrectly assume that two unrelated variables are related because they are similar. Such examples of these deceptive relationships appear in a variety of frameworks, including the disciplines of educational and behavioral psychology.

### **Implicit Theories**

Borman (1983, 1987) reported that performance ratings were affected by what he called “folk theories” of performance, that is, patterns that raters use to evaluate performance. The patterns are implicit theories that contain descriptions of what effective performance is, including which qualities an effective performance contains and

the relative importance of each trait. Borman (1987) further said one could make a case that implicit theories of performance comprise raters' unique definitions of performance. Similarly, Sternberg, Conway, Ketron, and Bernstein (1981) described implicit theories as individual interpretations of specific phenomena that exist in individual minds. Several other studies have examined how implicit theories held by raters influence the judgments they make while completing performance appraisals (Feldman, 1981; Hauenstein & Alexander, 1991; Ilgen & Feldman, 1983; Lord, Foti, & Phillips, 1982).

Dweck, a psychologist, and her associates studied educational settings to examine the influence of implicit theories (Dweck, Chiu, & Hong, 1995; Dweck & Leggett, 1988). They considered implicit theories and whether an individual views that he has a fixed way of thinking and judging the world, or if he is able to change his views after receiving new information or viewing different behaviors. More recently, Heslin, Latham, and VandeWalle (2005) examined the possible use of Dweck's implicit theories in industrial/organizational psychology. Research regarding implicit theory reported in the industrial/organizational psychology literature has studied measures aimed at understanding implicit theories in areas that were not originally considered. Lord and Maher (1993) offered that the way subordinates categorize and understand their leaders is partly based on the thinking that subordinates have regarding behavioral and character expectations. If an individual has in mind implicit theories of leadership, he or she may identify leadership behavior in others that match his or her implicit theory (Kenney, Schwartz-Kenney, & Blascovich, 1996).

### **Implicit Covariance Model**

To understand how sources think about and then rate participants, Woehr, Day, Arthur, and Bedeian (1996) formulated a systematic distortion hypothesis through which ratings are systematically distorted in the conceptual thinking of the raters. The authors' research suggested that such schemes are prone to be used in appraisals when raters are not knowledgeable about a particular job or when there is a time lapse between observing the performance and providing the ratings (p. 418). The systematic distortion process has also been linked to the "halo effect" in explaining source ratings.

### **General Impression Model**

According to Woehr et al. (1996), raters may possess an overall impression of a ratee's performance, causing systematic distortion. That impression would serve as the foundation for performance ratings (p. 419). Woehr et al. (1996) conjectured that raters often conceptualize behaviors into specific clusters and then arrive at a general opinion about the individual being rated. Thus, the rater has formed an overall impression of a ratee and stored it in memory. Subsequently, either the general impression or the specific information about behavior may serve as the source for performance ratings. The writers concluded that an overall impression is formed and utilized when raters complete performance evaluations. Woehr et al. (1996) found the following in concurrence: Lingle and Ostrom (1979); Skowronski and Carlston (1987); and Wyer and Srull (1989). Therefore, raters will use an overall impression in the performance appraisal process unless they have specific behavioral examples in mind (Woehr et al., 1996, p. 420).

Research has found other possible rater sources of bias in multisource feedback. For example, the common practice of ratees selecting their raters may be introducing new

biases (Thiry, 2009, p. 52). Feedback accuracy also depends on the rater's assurance of confidentiality, the rater's concerns and feelings regarding the feedback process, the opportunity the rater has had to see the ratee in action, the relationship of the rater and ratee, the length of time of their relationship, and whether the overall rating is positive. Ratings are also susceptible to leniency bias (Farh & Werbel, 1986; Farh, Werbel, & Bedeian, 1988). For example, Sala (2001) believed that the scores that ratees mark for themselves are influenced by the job level the ratee holds; that is, a higher job level will yield a higher score.

Alimo-Metcalf (1998) noted that a major area of growth in the field of leadership development is the use of 360 degree performance feedback (p. 35). She said that performance feedback enhances the growth of the ratee and provides the ratee with the tools to improve performance. One way to assure that development planning is effective is to implement a development process for all levels of the leadership team. Leadership development needs to evolve to a point where it is regarded as a way for an organization to complete its investment in the talent (Nahapiet & Ghoshal, 1998). Many contemporary organizations are striving to develop talent that relies on mutual trust to replace the traditional management structure of planning and control. The performance management process is essential to an organization's overall human capital development system and is one of the most useful tools for shaping personal behavior within an organization (Weatherly, 2004).

McCauley and Van Velsor (2004) connected assessment data, which can come from self evaluation or other sources, Assessment is important because, with guided reflection (Zentis, 2007), it gives people an understanding of how they are perceived,

which will help them develop their leadership style and recognize their primary development needs. Assessment data point to the holes between a person's self-awareness about current capabilities and performance and a realization of full potential. A gap analysis may conclude that an individual needs an action plan based on job requirements, career goals, others' expectations, or one's own expectations.

A significant benefit of multisource performance feedback is the self-appraisal of the ratee. Researchers have found that self-appraisals are worthwhile. For example, individuals can improve performance by realizing their belief in themselves (Bretz, Milkovich, & Read, 1992). Reflection on inconsistencies in ratings between self- and others can also provide a clear perception about the ratee's effectiveness (Jones & Bearley, 1996; McCauley & Moxley, 1996). The disparity between one's abilities and a desired state is one explanation of why self-reflection may inspire developmental planning. If a lack of a particular ability is important to the ratee, and if the ratee has confidence in the assessment, the ratee will work to reconcile the disparity by improving his or her abilities. If the assessment results show that the rated individual has a particular accomplishment or skill, then the assessment may increase the ratee's self-confidence. As a result, the person may look for new opportunities to build on current talents (McCauley & Van Velsor, 2004, pp. 6-7).

### **Conclusion**

Today, organizations desire to implement 360 degree or multisource performance feedback, both in local and global companies (Rowson, 1998). The literature shows that organizations continue to support multisource performance feedback as an assessment process for leadership development (Leslie & Van Velsor, 1995). However, rating

accuracy is often questioned. An organization must determine whether the process is appropriate in terms of the organization's strategy, and the individual being rated must accept the process as a positive experience for his or her leadership development and professional growth. The pool of raters must take the process seriously and provide their unique and timely perspective on the rated individual's competencies.

An examination of the literature reveals no firm conclusions about the effectiveness of performance assessment as a bridge to leadership development. However, the growing use of multisource or 360 degree feedback assessment indicates that many business decision-makers believe in its effectiveness. While this paper also includes research to the contrary, the literature in this review most often reflects that 360 degree or multisource feedback has value. Coupled with coaching and reflection, feedback from multiple rater sources can enhance the ratee's perceptions and help identify gaps in leadership competencies. There are many studies in other disciplines that have used the conceptual association or normative pattern matrix method to collect and compare data. However, as far as this researcher can determine, this study is unique in the field of multisource assessment in its use of conceptual association or normative pattern to provide external data to compare with actual data. The researcher interpreted the differences between rater sources and the normative correlation matrix.

For organizations to use multisource feedback as a leadership developmental tool, the participant and the organization need to be confident that the instrument is assessing the competencies stated. Comparing the patterns of a normative model with archival data provided data to examine what makes ratings from different rater sources different. This researcher found no studies that examined source differences by comparing a normative

correlation matrix pattern of ratings with a correlation matrix of actual ratings in multisource feedback. Therefore, the research project fills a void in the knowledge of the usefulness and accuracy of multisource feedback and can lead to better methods of administering leadership feedback performance assessments.

## CHAPTER 3: METHODOLOGY

### Introduction

The interpretation of performance assessment information can determine whether multisource feedback is perceived as a tool that improves organizational performance (Bradley, 2004, Abstract). In multisource feedback processes, individuals confront their perceived leadership competencies, based upon their own perspectives and those of individuals within their sphere of influence. Multisource performance reviews are believed valuable for understanding performance because the results provide different raters' assessments from various organizational perspectives (Borman, 1997). The 360 degree performance feedback is one such widely used instrument.

Accuracy of the feedback instrument is critical to the ratee's acceptance of the results and its value as a developmental tool. Multisource performance reviews continue to be popular for determining ratees' skills and developing activities aimed at performance improvement. "Still, important questions remain as to the psychometric properties of these popular tools, not the least of which is to what extent should we be concerned about systematic differences observed in different raters' perspectives of target performance?" (Lance, Hoffman, Gentry, & Baranik, 2008, p. 230).

By comparing archival data from an international company's multisource feedback process and a normative pattern derived from a questionnaire administered to and completed by a group of senior leaders of a second international company, this researcher examined source differences and the influence of implicit theories on answers on the assessment instrument, rather than true assessments from the raters' experiences



with the subject and observations of the subject's behavior. Understanding and building upon rater perspective may result in improved assessments (Bradley, 2004, Abstract).

This study posed the following research question, "What is the nature of the differences between performance ratings from different rating sources?"

1. Are the differences between the sources' correlation matrices for the actual data statistically significant?
2. Are the differences between the sources' correlation matrices for the actual data source types vs. the correlations for the normative group statistically significant?
3. What are the relative magnitudes of the differences between the correlation matrices from different sources?
4. How much difference do implicit performance theories make in the total rating scores that ratees receive?
5. How do the structures of the matrices for various sources differ?
6. How similar is the rotated factor structure (of the maximum number of non-error principal components) to the similarly rotated factor structure of the normative group?

### **Research Design**

Research design refers to the plan that links the methodology to the specific research methods (Crotty, 1998). This study utilized a nonexperimental quantitative research design to determine a comparative relationship of one correlation matrix to another (normative versus archival data) by examining differences between the matrices and source differences. To conduct this study, two data sets were gathered and

compared: The researcher compiled a set of archival data from EADS' multisource feedback assessment campaigns from 2006 to 2009 and a set of conceptual ratings based on the answers to a questionnaire describing the same competencies used in the EADS multisource feedback assessments as shown in Figure 2. The researcher utilized SPSS 17.0 to investigate the comparative relationships. Kerlinger (1986) wrote:

*Nonexperimental research is systematic empirical inquiry in which the scientist does not have direct control of independent variables because their manifestations have already occurred or because they are inherently not manipulable. Inferences about relations among variables are made, without direct intervention, from concomitant variation of independent and dependent variables [source's italics]. (p. 348)*

This study was quantitative in that the research referred to inquiries in which numerical values were assigned to some or all of the objects under investigation and in which the analysis used to draw conclusions was based on gathering numerical data and analyzing and interpreting the information (Tatsuoka & Silver, 1988, p. 677). The study was correlational in that it investigated a relationship between the two correlational matrices (Mertens, 2005). This research methodology is descriptive, cross-sectional, and nonexperimental in nature, using a set of *ex post facto* archival data from EADS and a set of conceptual ratings based on answers to a questionnaire describing the same competencies used in multisource feedback assessments (Bordens & Abbott, 2008). *Ex post facto* was most appropriate because the data are historical in nature.

To decide whether the main objective of the research is descriptive, one should answer the following questions:

- a. Were the researchers primarily describing the phenomenon?
- b. Were the researchers documenting the characteristics of the phenomenon? If the answer is “yes” (and there is no manipulation), then the term *descriptive nonexperimental research* should be applied.

In cross-sectional research the data are collected from research participants at a moment in time or during one relatively brief period (called contemporaneous measurement). The data directly apply to each case at that moment in time, and comparisons are made across the variables. Both research objective and time dimensions may be used to determine how to design the planned nonexperimental research (Johnson & Christensen, 2000).

### **Target Population and Sample**

Triola (2007) defined the target population to be “the complete collection (including all subjects) of scores, people, measurements to be studied” (p. 4). The population for the actual performance data for this study came from an archived dataset of a global company, EADS, that included anonymous self-ratings of 1,210 business leaders from six business units, and more than 10,000 other raters who completed the 360 degree feedback assessment in a single point of time between 2006, the year of its inception in the company, and March 2009. “Other raters” comprised 1,395 higher level managers, 5,233 peers, 4,964 direct reports, and 4,056 customer/clients. It is possible to measure the entire population with the support of the company’s Information Technology organization.

There are advantages to using archival data: Archival data add to presumptions from earlier events, assist users in assessing interventions, provide a basis for further

study, and address questions within specific disciplines (Funder, Parke, Tomlinson-Keasey, & Widaman, 1993). Disadvantages include the possibility of using outdated methodology compared with data collected using modern techniques and the potential problem of missing data.

The sample for the conceptual normative data came from a survey completed by 32 senior leaders of Force Protection, Inc., a multinational defense company. The survey asked them to rate six leadership competencies (Figure 2). In non-probability sampling, the comparison of the sample to the population remains unknown (McDaniel & Gates, 1991). The non-probability judgment sample of 32 senior leaders is consistent with the sample sizes used in previous studies in which “judges” rated hypothetical data in order to form a normative data set (Bales, 1970; Dahlstrom & Welsh, 1960; Newcomb, 1931; Sears, Macoby, & Levin, 1957; Shweder, 1975.)

The compiled results provided a group perspective on the ratings of the six competencies. Those ratings became the normative or conceptual model for the research project. The conceptual correlations between the competencies can be operationalized by use of the questionnaire in Appendices A and B. Specifically, one-half the participants answered the question: “Of 100 people, or workers, who are above average in ‘A,’ how many would you expect also to be above average in ‘B’?” Then, the second half of the responders answered the reverse of A and B.

### **Instrument**

The EADS instrument is similar to and derived from the PROFILOR®, which has been found to be both valid and reliable (Lin, 2007, p. 86). PROFILOR®, developed by

Personnel Decision International Corporation, is meant to characterize competencies generally requisite for managers in a changing business environment (PDI Web site, 2011).

The development of the PROFILOR® was a rigorous process. The instrument is based on several decades of consulting experience and research on management, including an extensive review of the management and psychology literature, a thorough analysis of the large Management Skills Profile data base that included more than 20,000 managers, comprehensive job analysis, and group interviews with hundreds of managers from many functional areas and most major industries, and pilot testing. (PDI Web site, 2011)

The EADS instrument is a customized modification of the PROFILOR®. The instrument is proprietary. It is designed to identify leadership behaviors and traits that the corporation wants and expects from its leaders. The purpose of the instrument is to collect feedback from participants to determine the developmental needs of the leader. It is made up of six higher order dimensions and 16 leadership competencies. EADS has defined each of the six core competencies as shown in Figure 2. Raters use a 6-point Likert-type scale for assessment of observed behaviors. The raters have the following choices: N/A, meaning not observed; never; rarely; sometimes; frequently; and always. The answers are rated on a scale ranging from 1 = never to 5 = always. N/A has the value of zero. The multisource process is administered online and the raters are guaranteed confidentiality of their responses. Participants are emailed multisource feedback assessments to be completed by themselves as well as subordinates, peers, managers, and customer/clients.

Participants are allowed to choose which of their coworkers/customers/clients will complete the performance ratings. The completed surveys are emailed by each rater to EADS. EADS sends the completed forms to Technomedia, a company hired to code, enter, check, and prepare reports of the data from participants and their selected reviewers. As is the norm in the 360 degree assessment process, each participant receives a comprehensive report that shows self-ratings compared to peer, boss, subordinate, and customer/client ratings. The participant then schedules a facilitator session with one of the 200 certified facilitators the corporation employs. The participant, assisted by a facilitator or an executive coach, may construct a developmental action plan based on feedback from the report. Participants are encouraged to meet with their manager and human resources representative to discuss their results and development plan.

**Conceptual Questionnaire.** The conceptual correlations between the competencies can be operationalized by use of the questionnaire shown in Appendices A and B. The form used to obtain the conceptual ratings consisted of the same six competencies shown in Figure 2 that make up the EADS multisource feedback assessment instrument.

### **Validity and Reliability**

For a research process or result to be described as reliable and valid means that it has gone a long way toward gaining scientific acceptance (Carmines & Zeller, 1979, p. 15). Van Velsor, Leslie, and Fleenor (1997) described reliability and validity testing as part of the instrument development process. Development, they posited, generally begins with listing behaviors or characteristics of effective management or leadership (p. 6).

Ultimately, they wrote:

The quality of the final product depends on (a) the combined quality of the theory, research, and experience of the developer; (b) the developer's skill in translating theory, research, and experience into written items; and (c) the developer's attentiveness to instrument development and feedback design. If individuals are left to process feedback (no trainer or facilitator is available), or if an instrument is not accompanied by comprehensive interpretive and development materials, the instrument's content must be extremely clear. The harder the items are to interpret, the more difficulty managers will have in benefiting from the feedback and the more important the quantity and quality of support becomes. (Van Velsor et al., 1997, pp. 7-8)

Typically, developing scales for the instrument involves using both statistical factors and rational/intuitive judgment (Van Velsor et al., 1997, p. 9). The statistical approach can involve grouping items together and forming scales that represent similar rating schemes. Recommended statistical process can be factor analysis, item-scale correlations, and cluster analysis (Van Velsor et al., 1997, p. 9). The statistical processes can involve grouping items into scales based on the degree of similarity in response to patterns of the raters; that is, factor analysis, cluster analysis, or item-scale correlations. The rational/intuitive approach uses the author's judgment and experience about which items should be grouped together to form a scale. Instrument development may involve either process or both. Grouping items on scales according to judgments involving how the competencies relate to each other and eliminating items that do not fit operationally provides a more effective instrument (Van Velsor et al., 1997, p. 9).

Although the EADS 360 degree feedback questionnaire was based on the PROFILOR® and the PROFILOR® is regarded as valid and reliable (Lin, 2007, p. 86), validity and reliability studies were conducted on the customized EADS 360 Feedback System. The Cronbach's alphas scores for the PROFILOR® indicate a median of 0.89 across rater sources, indicating the PROFILOR® multisource assessment to be a reliable instrument with acceptable mean Cronbach's alpha value above 0.70. The interclass correlation coefficients range from 0.47 to 0.60 for peers and 0.48 to 0.61 for direct reports, higher than the optimal inter-item correlation values range of 0.2 to 0.4 (as recommended by Briggs & Cheek, 1986) and indicating the PROFILOR® multisource assessment to be a reliable instrument (Personnel Decisions International, 2000).

**Validity.** Content validity relates to the degree to which the items represent a comprehensive measure of the leadership competencies in question. A description of the scale's purpose should guide the process of crafting or selecting items for the measurement (Hooper, 2004, p. 46). For example, a 360 degree feedback instrument should cover areas for effective performance for the specific job (Van Velsor et al., 1997, p. 19). Often, the information can be gleaned from the theory or model on which the instrument was based or from the experience of the developer. Content involves experience and judgment in addition to statistical analysis. However, Van Velsor et al. (1997) said that it is important to remember "there is no instrument that is perfect for everyone, no instrument that will always be ideal, and no instrument that cannot be improved" (p 19).

Validity is a somewhat theoretical issue, according to Rungtusanatham (1998), because it raises the question, "Valid for what purpose? Is the instrument measuring



what its sponsors think it is measuring?” (p. 11). To determine the validity of 360 degree feedback instruments one must first examine whether the instrument measures what it is intended to measure. For example, if leadership competencies are important to the organization, does it measure those competencies? One must then determine whether what is measured will make sense. That is, are higher scores actually a determinant of organizational effectiveness? The third step is evaluating whether feedback, leadership development, or action learning would influence what is measured.. Then, Van Velsor et al. (1997) asked, “Do the instrument’s scales represent domains that are amenable to development?” (p. 13).

In her dissertation on developing a 360 degree instrument with reliability and validity, Hooper (2004) listed the first step as establishing an instrument’s purpose and intended use. Defining and limiting the scope of the instrument creates the foundation for further development. Hooper then discussed the second step, which builds content validity into the design process. Recognizing and defining the theoretical constructs will further clarify and limit the boundaries of the instrument, she wrote (p. 45). She believed it necessary to understand the complexity of the subject matter at the construct level so that the scales could address the major dimensions of the subject matter. Each scale should seek to address the scope and complexity of the construct it is intended to address. Operational definitions for each construct in the theoretical framework must be completed before abstract constructs are measurable (Hooper, 2004, p. 45).

In the development of an instrument, validity is reality testing. Validity testing is where the developer begins to know whether the model is meaningful for real managers in real jobs. In other words, validity is integrity. If the developer does not know whether

an instrument does what it says it will do, and, if there is no evidence that scores are related to job effectiveness, the facilitator/coach will not understand how to interpret that instrument's data (Van Velsor et al., 1997, p. 13). According to Nunnally and Bernstein (1994), the two standards for ensuring content validity are (a) the instrument should have a representative collection of items, and (b) the instrument should be constructed using sensible methods. For performance assessment, items in the assessment instrument may include the more important traits, behaviors, and performance outcomes attributable to the ratee (Latham & Wexley, 1994).

In the process of developing the EADS instrument, a comprehensive set of items was analyzed using sensible methods (Nunnally & Bernstein, 1994). Gröesch (N. Gröesch personal communication, September 15, 2009) described the steps she used in developing the instrument and ensuring validity:

Step 1: Designed leadership competency model

Step 2: Sorted leadership competencies into potential items for the 360 feedback instrument; those were behavioral, task-based, and skills based

Step 3: Drafted items similar to PROFILOR® and validated content against EADS' leadership competencies. Content validity evidence for the EADS instrument was obtained by ensuring that the content of the data collection instruments adequately incorporated the theories and concepts associated with the content domain of interest (Cronbach & Meehl, 1955; Lawshe, 1975; Shepard, 1993).

Step 4: Tested items internally for readability and clarity, then translated and back-translated into the languages of EADS employees. Gröesch convened a focus group of individuals with acknowledged expertise in the subject matter under examination to

judge whether items covered all aspects of the domain that were intended to be measured and whether they were in appropriate proportions relative to that domain as advocated by Gliner, Morgan and Leech (2009, p. 159). In writing and orally, the focus group reviewed and provided feedback on clarity, relevance, and unintended threat in the items, as well as ethnic and gender bias.

Carmines and Zeller (1979) argued that reliability and validity of the instrument must be proven in order for the data collected and analyzed to be useful for interpretation of results. Conversely, they said that when the measurements are not valid and reliable, the data analysis performed may lead to erroneous conclusions. If measurements are not reliable or valid, analysis of the data could lead to improper assumptions and incorrect conclusions about the underlying constructs. EADS used competencies from subject matter experts who had substantial knowledge and experience with leadership and performance measures needed for a multisource feedback assessment. Those individuals helped operationalize the primary domains of interest and, thus, supported construct validity. Construct validity evidence for the EADS instrument was validated by analyzing threats such as inadequate operationalization of constructs, for example, inaccurate or incomplete definition of constructs, dimensions, variables, and measures (Shadish, Cook, & Campbell, 2002).

Step 5: Sent items to external provider to load online as the EADS 360 feedback instrument. Gröesch concluded that the degree of agreement between content expert reviewers and the author served as a check of the construct validity and the reliability of the instrument. In another context, Hooper (2004, p. 52) concurred. The EADS instrument is evaluated periodically for possible revisions.

Singleton, Straits, and Straits (1993) held that “validity cannot be assessed directly” (p. 121). DeVellis (1991) indicated that validity can “only be inferred from the manner in which [a measurement instrument] was constructed [i.e., *content validity*], its ability to predict specific events [i.e., *criterion-related validity*], or its relationships to measures of the same and different constructs [i.e., *construct validity*]” (p. 43). He further wrote that validity draws attention to the “extent of matching, congruence, or ‘goodness of fit’ between an operational definition and the [construct] it is purported to measure” (p. 115). The instrument’s validity, then, matches an assessment of the instrument as it fits operationally for a specific concept (DeVellis, 1991, p. 43).

Rungtusanatham (1998) wrote that the construct validity of an instrument is determined by the items representing the unique facts of a construct (p. 11). It is the extent to which a measurement instrument reflects the parts of a construct. Both concurrent and predictive validity relate the scores to other attributes. They are not inherent to the instrument, and both are temporary in nature. Concurrent validity examines the instrument to determine the relationship between ratee’s scores and performance measured at the same time (Van Velsor et al., 1997, p. 17).

Researchers obtain evidence of *criterion validity* by comparing the scores obtained from the instrument under evaluation to the scores on a second test or procedure, referred to as the criterion, that is presumed to measure the same variable (Fraenkel & Wallen, 2006). Based on the definition of criterion-related validity, its application goes beyond the span of this study. However, it may be possible to incorporate criterion-related validity studies in subsequent re-evaluations of the EADS instrument for validity.

**Reliability.** A prerequisite for having *valid* data is having *reliable* data.

*Reliability* is the “consistency of repeated measurements across persons” (Carmines & Zeller, 1979, p. 31). Fraenkel and Wallen (2006) expanded on that definition: “Reliability refers to the consistency of scores—how consistent they are for each individual from one administration to another and from one set of items to another” (p. 157). Possibilities for reliability testing are test-retest reliability, parallel forms, or alternate forms reliability. The test-retest method is a comparison of results from an initial test with follow-on repeated measures. If the instrument is reliable and if variables remain unchanged, it is expected that there will be close agreement over repeated tests. EADS uses the instrument only once for developmental purposes of the ratee.

Latham and Wexley (1994) defined internal consistency in practical terms as “an indication of the homogeneity or ‘sameness’ of the items [on] a scale” in an assessment instrument (p. 68). A commonly used test for internal consistency is Cronbach’s alpha because of its ease of use and its incorporation of positive aspects of other reliability techniques (Carmines & Zeller, 1979). That approach is based upon item covariance for measuring the reliability of the EADS instrument, that is, the test for internal consistency utilizing Cronbach’s alpha calculations. Cronbach's alpha ( $\alpha$ ) is a commonly used method in determining the reliability of scores (Gall et al., 2003). The following formula is applied:

$$\hat{\alpha} = \frac{k}{k-1} \left( 1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_x^2} \right)$$

Alpha is determined by calculating the average correlation of each item of a scale with every other item. Cronbach (2004) noted that the alpha coefficient was developed for

measuring differences among individuals but now is more widely used for reliability analysis.

In behavioral science research and literature, Cronbach's alpha is a preferred method for reliability testing. As with other reliability coefficients, alpha should be above 0.70; however, journal articles are common where one or more scales are reported to have somewhat lower alphas, that is, in the 0.60-0.69 range, especially if the scale has only a small number of items (Cronbach, 2004, p. 416). A very high alpha, that is, greater than 0.90, may indicate that there is repetition and redundancy in the scale.

Cronbach's coefficient alpha may provide a measure of reliability to an administration of an instrument or questionnaire. These reliabilities for the EADS multisource feedback instrument were computed by first computing the mean rating for each rater type for each ratee on each of the 70 behavioral items. Thus, for each of the 16 competencies for each ratee, there could be a maximum of six mean rater type scores. The unweighted mean across all rater type mean scores was then computed for each behavioral item. That process resulted in one multirater summary score for each behavioral item for each ratee. Those were the scores used in calculating the overall reliabilities of the 16 competency scales.

Each of the 16 EADS multisource assessment competencies for each ratee consisted of a maximum of six mean rater type scores measuring internal consistency reliability (Cronbach's alpha) from 0.59 to 0.81 (mean 0.73) indicating the EADS multisource assessment to be a reliable instrument with acceptable mean Cronbach's alpha value above 0.70. An additional reliability analysis was performed using inter-item correlations. The mean inter-item correlations were in the 0.4 to 0.6 range, higher than

the optimal inter-item correlation values range from 0.2 to 0.4 as recommended by Briggs & Cheek, 1986 indicating the EADS multisource assessment to be a reliable instrument. Experts like Rungtusanatham (1998) pointed to the reason that validity and reliability of the EADS 360 instrument matters: Only when instruments are valid and reliable can an analysis of the data assure the participants that the results may be useful for their intended purpose.

### **Data Collection**

EADS provided the archived data used with permission for this study and generated historical aggregated mean ratings of each rater category for a population of more than one thousand 360 degree feedback events. EADS removed participants' names and assigned a numerical code that protected the anonymity of the participants. The company electronically sent the coded data to this researcher in an Excel file. Data were organized for statistical analyses using SPSS 17.0 software. The researcher was responsible for all data analysis and other data interpretation/presentation work.

The researcher administered the questionnaire shown in Appendices A and B to the 32 Force Protection senior leaders. First, the researcher met with the chief operating officer of Force Protection to explain the study, to obtain his support, and to request and obtain a letter of approval, included as Appendix C. Then, a letter was emailed to him detailing again the purpose of the study and requesting a date to administer the survey. The researcher asked him to invite volunteers to participate in the research project. Participants received a statement to explain how their anonymity would be maintained and to state that the survey information would be used only for the study and not viewed by anyone other than the researcher. The researcher provided instructions and ensured

confidentiality. Participants completed a consent form and the questionnaire and returned them to the researcher. After the individual data were collected from the brief questionnaire, the researcher downloaded the results from the survey to SPSS 17. The researcher generated reports from the SPSS 17 and analyzed the results. The final data will be stored for seven years for future research. The data from SPSS will be stored for up to seven years and then discarded.

### **Data Analysis Plan**

**Correlation Matrix of Archival Data.** The researcher received the archival data on an Excel spreadsheet. The researcher exported the raw data into SPSS and screened it for accuracy, completeness, and suitability. Raw data were examined to ensure that data fields were complete and the values were within expected ranges. After confirming data quality, the descriptive statistics and graphical plots were generated to gain additional information on the quality and characteristics of the data.

**Correlation Matrix of the Normative Pattern.** After receiving the data from questionnaires administered at Force Protection as shown in Appendices A and B, the researcher followed the steps described to form the normative correlation matrix: When asking, “Out of 100 randomly selected people who performed above average on Competency A, how many would you expect to have performed above average also on Competency B?” (half if A, then B; half if B, then A). If it is assumed that both competencies are normally distributed, then by definition there will also be 100 people who are below average on the two competencies. From this stipulated group of 200 people, that is, 100 above average on each competency and 100 below average on each competency, infer a 2x2 table as shown in Table 1 with the following marginals:



Table 1

*2x2 Table of Competency Ratings*

Competency A	Competency B		Totals
	Below average	Above average	
Above average	A	B	100
Below average	C	D	100
Totals	100	100	200

The estimate requested in each rating was for the cell marked B in Table 1. The 32 answers given for each competency pair (half if A then B, half if B then A) were averaged to produce a single estimate for cell B, for example, 40, for the respective pair of competencies. From this estimate, the frequencies for cells A, for example, 60; C, for example, 40; and D, for example, 60, can be computed by subtraction. The information on the 2x2 table can then be converted to a tetrachoric correlation, which is approximated by the following formula:

$$r_{tet} = \cos (180/(1 + \text{sqrt}(BC/AD)))$$

**Tetrachoric Correlation.** The tetrachoric correlation is computed only for, or applicable to, 2x2 tables. The tetrachoric correlation coefficient estimates the correlation between the pair of variables. The tetrachoric correlation estimates what the underlying Pearson correlation would be between two normally distributed, continuously measured variables that have been artificially dichotomized, for example, at the average, and represented in a 2x2 table. This procedure was followed for all 15 unique pairs of the six competencies to produce the normative correlation matrix.

The actual competency correlation matrix of the archival multisource performance rating data for each source was compared to the correlation matrix of the

conceptual data to determine, “What is the nature of the differences between performance ratings from various rating sources?” “The only way to test a hypothesis is to eliminate alternatives of the hypothesis” (Anderson, 1966, p. 9). The researcher analyzed the data and examine these research hypotheses by seeking answers to the following research questions:

1. Are the differences between the sources’ correlation matrices for the actual data statistically significant?

H<sub>1</sub>: There is significant difference between the correlation matrices for self vs. manager.

H<sub>2</sub>: There is significant difference between the correlation matrices for self vs. direct report.

H<sub>3</sub>: There is significant difference between the correlation matrices for self vs. peers.

H<sub>4</sub>: There is significant difference between the correlation matrices for self vs. customer/clients.

H<sub>5</sub>: There is significant difference between the correlation matrices for direct reports vs. peers.

H<sub>6</sub>: There is significant difference between the correlation matrices for peers vs. customer/clients.

H<sub>7</sub>: There is significant difference between the correlation matrices for manager vs. customer/clients.

H<sub>8</sub>: There is significant difference between the correlation matrices for direct reports vs. peers.

H<sub>9</sub>: There is significant difference between the correlation matrices for direct reports vs. customer/clients.

H<sub>10</sub>: There is significant difference between the correlation matrices for peers vs. customer/clients.

This research question addressed whether the differences between the sources' correlation matrices are statistically significant. The pairs of matrices for each combination of the five source types, self, subordinate, peer, manager, and customer/client, were compared using the asymptotic  $\chi^2$  test proposed by Jennrich (1970).

2. Are the differences between the sources' correlation matrices for the actual data source types vs. the correlations for the normative group statistically significant?

H<sub>11</sub>: There is significant difference between the correlation matrix for self vs. the correlations for the normative group.

H<sub>12</sub>: There is significant difference between the correlation matrix for direct reports vs. the correlations for the normative group.

H<sub>13</sub>: There is significant difference between the correlation matrix for managers vs. the correlations for the normative group.

H<sub>14</sub>: There is significant difference between the correlation matrix for peers vs. the correlations for the normative group.

H<sub>15</sub>: There is significant difference between the correlation matrix for customer/clients vs. the correlations for the normative group.

This question was addressed by asking whether the differences between the normative correlation matrix derived from the questionnaire of perceived covariation and the correlation matrices of the five source types are statistically significant. The normative correlation matrix was compared to each source type matrix using the

asymptotic  $\chi^2$  test proposed by Jennrich (1970).

3. What are the relative magnitudes of the differences between the correlation matrices from different sources?
  - 3a. What is the relative magnitude of difference between the correlation matrices for self vs. manager?
  - 3b. What is the relative magnitude of difference between the correlation matrices for self vs. direct reports?
  - 3c. What is the relative magnitude of difference between the correlation matrices for self vs. peers?
  - 3d. What is the relative magnitude of difference between correlation matrices for self vs. customer/client?
  - 3e. What is the relative magnitude of difference between the correlation matrices for manager vs. direct reports?
  - 3f. What is the relative magnitude of difference between the correlation matrices for manager vs. peers?
  - 3g. What is the relative magnitude of difference between the correlation matrices for manager vs. customer/client?
  - 3h. What is the relative magnitude of difference between correlation matrices for direct reports vs. peers?
  - 3i. What is the relative magnitude of difference between the correlation matrices for direct reports vs. customer/clients?
  - 3j. What is the relative magnitude of difference between the correlation matrices for peers vs. customer/clients?

This question was addressed by computing the Euclidean distance between each pair of correlation matrices. These distance scores do not follow a probability distribution and cannot be tested for significance. Their utility in the present study was purely as a basis for making quantitative comparisons between matrices of their similarity to a target matrix.

4. How much difference do implicit performance theories make in the total rating scores that ratees receive?

4a. Are there differences in the correlations between weighted mean scores for source types and the predicted scores based on source types' regression equation?

H<sub>16</sub>: There is a difference in the correlations between weighted mean scores for source type self and the predicted scores based on source type-managers' regression equation.

H<sub>17</sub>: There is a difference in the correlations between weighted mean scores for source type-self and the predicted scores based on source type-direct reports' regression equation.

H<sub>18</sub>: There is a difference in the correlations between weighted mean scores for source type-self and the predicted scores based on source type-peers' regression equation.

H<sub>19</sub>: There is a difference in the correlations between weighted mean scores for source type-self and the predicted scores based on source type-customer/clients' regression equation.

- H<sub>20</sub>: There is a difference in the correlations between weighted mean scores for source type-manager and the predicted scores based on source type-direct reports' regression equation.
- H<sub>21</sub>: There is a difference in the correlations between weighted mean scores for source type- manager and the predicted scores based on source type-peers' regression equation.
- H<sub>22</sub>: There is a difference in the correlations between weighted mean scores for source type-manager and the predicted scores based on source type-customer/clients' regression equation.
- H<sub>23</sub>: There is a difference in the correlations between weighted mean scores for source type-direct reports' and the predicted scores based on source type-peers' regression equation.
- H<sub>24</sub>: There is a difference in the correlations between weighted mean scores for source type-direct reports' and the predicted scores based on source type-customer/clients' regression equation.
- H<sub>25</sub>: There is a difference in the correlations between weighted mean scores for source type-peer and the predicted scores based on source type-customer/clients' regression equation.

The tests of the differences in means were conducted using the paired t-test and results reported in Chapter 4.

- 4b. Are there differences in the means between actual total scores for source types and the predicted scores based on source types' regression equation?

- H<sub>26</sub>: There is a difference in the means between actual total scores for source type-self and the predicted scores based on source type-managers' regression equation.
- H<sub>27</sub>: There is a difference in the means between actual total scores for source type-self and the predicted scores based on source type-direct reports' regression equation.
- H<sub>28</sub>: There is a difference in the means between actual total scores for source type-self and the predicted scores based on source type-peers' regression equation.
- H<sub>29</sub>: There is a difference in the means between actual total scores for source type self and the predicted scores based on source type-customer/clients' regression equation.
- H<sub>30</sub>: There is a difference in the means between actual total scores for source type-manager and the predicted scores based on source type-direct reports' regression equation.
- H<sub>31</sub>: There is a difference in the means between actual total scores for source type-manager and the predicted scores based on source type-peers' regression equation.
- H<sub>32</sub>: There is a difference in the means between actual total scores for source type-manager and the predicted scores based on source type-customer/clients' regression equation.
- H<sub>33</sub>: There is a difference in the means between actual total scores for source type-direct reports and the predicted scores based on source type-peers' regression equation.
- H<sub>34</sub>: There is a difference in the means between actual total scores for source type-direct reports and the predicted scores based on source type-customer/clients' regression equation.
- H<sub>35</sub>: There is a difference in the means between actual total scores for source type-peer and the predicted scores based on source type-customer/clients' regression equation.

The sum of each rater's ratings was regressed on the six competencies separately for each source type. Multiple regression is a correlational research method that enables a researcher to determine what traits and behaviors, the variables, are the best predictors of criterion variables, that is, performance outcomes (Fraenkel & Wallen, 2006).

The resulting regression equations expressed the relative degree to which the respective source type differentiated rates on the six competencies, hence, the relative influence the competencies had on the source type's judgment and their inter-dependencies within each source type's cognitive framework. Each source type's regression equation was used to compute predicted scores for each other source type's ratings. The correlations between the predicted ratings and each source type's weighted mean scores were computed as an indicator of how much of a difference was caused by the implicit performance theories used by each source type. The tests of the differences in mean were conducted using the paired t-test.

##### 5. How do the structures of the matrices for various sources differ?

This research question was addressed by first extracting the maximum number of non-error factors from a principal components analysis (PCA) of the six competency variables performed separately for each source group. The PCA *components matrix* in SPSS® lists the unrotated loadings on the variables and the *rotated component matrix* lists the variable loadings from orthogonal, that is, Varimax, rotations. Varimax rotations are called orthogonal rotations because the axes that are rotated remain at right angles to each other. Varimax rotation was performed on each resulting set of factors.

This research question inquired as to the comparability between source types of the composition of the optimum number of rotated principal components computed for



each source type's ratings of the six competency dimensions. These comparisons consisted of pairing the rotated factors that are most similar between source types in terms of the competency dimensions that define them and then computing the correlation between the rotated factor loadings for each pair of factors. Essentially, this was a descriptive qualitative comparison to reveal the nature of any differences between the groups in how they define the more important factors they use for differentiating the rates.

6. How similar is the rotated factor structure (of the maximum number of non-error principal components) to the similarly rotated factor structure of the normative group?

The score for each ratee on each competency was averaged across all rating sources. The resulting averaged scores were then subjected to principal components analysis (PCA), and the maximum number of non-error factors were extracted. These were then be subjected to Varimax rotation. The comparisons consisted of pairing up the rotated factors that were most similar between the normative factor solution and that of each source type in terms of the competency dimensions that defined them and then computing the correlation between the rotated factor loadings for each pair of factors. Again, this was essentially a descriptive qualitative comparison to reveal whether the combined data for all the rating source types more closely corresponded to the factor structure of the normative group than any source type's structure did on its own.

### **Ethical Considerations**

This study addressed the ethical considerations in using human subjects as part of the study. The protection of participants' anonymity and confidentiality in this study was

ensured by the removal of all identifying information from the data prior to analysis. This study used both archival data and collected data through the participation of voluntary subjects who completed a questionnaire that asked for hypothetical answers with no penalty for withdrawing. Participants received information explaining the purpose and procedures for the research, along with a participation request. The questionnaire did not collect any personal identification data, and no information beyond the questionnaire data was given to the researcher, thus ensuring against risk of invasion of privacy and breach of confidentiality.

The results from the conceptual questionnaire were compiled in May 2011. The researcher found no identified risk related to situation, time, or population associated with either instrument. Prior to conducting the survey, the Dallas Baptist University Protection of Human Subjects form was completed and submitted (Appendix D). The researcher received approval from the Dallas Baptist University Committee for Protection of Human Subjects. The authorization is valid from 05/09/11 until 05/09/2018. The signed forms and completed survey responses secured by the researcher will be retained for seven years for future research.

The research met the ethical criteria described by Bogdan and Bilken (2003). These considerations include making certain there is informed consent between the researcher and participant in the study. Also, the study must be a voluntary activity for the participant. Participants are ensured that their privacy is protected. Finally, the participant shows respect to the participants and completes the process by reporting the results of the study with professional integrity (p. 43-44).

This chapter presented the research methodology, purpose and design of the study, and research questions. The target populations were presented first, followed by the multisource instrumentation, the hypothetical questionnaire development, and the validity and reliability of the instrument. The chapter concluded with a description of the data collection, data analysis, and ethical considerations.

## CHAPTER 4: RESULTS

Multisource feedback is capable of providing a more comprehensive understanding of individual performance than can be obtained through supervisor ratings alone (Hedge, Borman, & Birkeland, 2001). Supervisors rarely know all aspects of subordinate job performance; therefore, relying exclusively upon supervisor ratings would neglect aspects of job performance that are less visible to an immediate supervisor. The perspective provided by each rating source provides valuable insight into a ratee's performance (Hazucha, Hezlett, & Schneider, 1993). Ignoring the unique perspective of differing rating sources would exclude important job-relevant sources of information and settle for performance measures that are construct deficient. Such an alternative would be undesirable (Waldschmidt, 2006, p. 21).

Yet, although multisource feedback is used by numerous organizations to assess performance, questions remain about its efficacy. This study was designed to examine one aspect of multisource feedback ratings, that is, to compare the relationship between the normative pattern and archival pattern of ratings and to explore the possible interpretation for the patterns of ratings. By comparing actual archival data from one international company's multisource feedback process with a normative pattern derived from a questionnaire completed by a group of senior leaders of a second international company, this researcher examined the relationship between two correlation matrices and explored the possible interpretations for the patterns of ratings.

The study answered a series of six research questions through the development and examination of 35 relevant research hypotheses and 10 sub-research questions. This

chapter reports the data analysis and results of the statistical testing of the research hypotheses, followed by an analysis of the results.

## **Data Gathering, Analysis, and Results**

### **Data Collection**

EADS provided the archived data, used with permission, for a population of more than one thousand 360 degree feedback events. Technomedia is the firm that EADS hired to code, enter, check, and prepare the reports of the data submitted by participants and their selected raters. EADS removed participants' names and assigned a numerical code to protect the confidentiality of the participants. The company electronically sent the coded data to this researcher in an Excel file. Data were organized for statistical analyses using SPSS/PASW 17.0 computer software.

The researcher administered a questionnaire to 32 Force Protection senior leaders. The Chief Operating Officer of Force Protection invited participants to take part voluntarily in the research project. Participants received a statement to explain how their anonymity would be maintained and to convey that the questionnaire information would be used only for the study purposes. Participant anonymity was ensured as the questionnaire did not collect any personal identification data, and no information beyond the questionnaire data was passed to the researcher or to any third party. Participants completed a consent form and the questionnaire and returned them to the researcher. Data were collected from each brief questionnaire and entered into SPSS 17.0.

The results of the study are organized as follows: (a) description of the sample, (b) descriptive statistics, (c) preliminary data analysis, (d) details of analysis, and (e)

conclusions that summarize the results in relation to the study's research questions. Only complete questionnaires were utilized for the study.

### **Description of the Sample**

**Archival data.** The study accepted the archived dataset of EADS that includes anonymous self-ratings of 1,210 business leaders from six business units, and more than 10,000 other raters who completed the 360 degree feedback assessment between 2006, the year of its inception in the company, and March 2009. "Other raters" comprise 1,395 higher level managers, 5,233 peers, 4,964 direct reports, and 4,056 customers.

**Normative data structure.** The Chief Operating Officer of Force Protection sent email invitations to 34 company executives, with 32 participants completing the questionnaire shown in Appendices A and B in person on May 10, 2011. The study analyzed the 32 completed questionnaires to form the normative structure data. The non-probability sample of senior leaders was selected on judgment, using the leadership team of Force Protection. The sample is consistent with the sample sizes in other studies that used "judges" to rate hypothetical data in order to form a normative data set (Bales, 1970; Dahlstrom & Welsh, 1960; Newcomb, 1931; Sears, Macoby, & Levin, 1957; Shweder, 1975).

**Foundational Data.** The actual competency correlation matrix of the archival multisource performance rating data for each source was compared to the correlation matrix of the conceptual data to determine, "What is the nature of the differences between performance ratings from various rating sources?"

**Correlation matrix of the normative pattern.** The tetrachoric correlation estimated what the underlying Pearson correlation would be between two normally

distributed, continuously measured variables that have been artificially dichotomized, that is, at the average and represented in a 2x2 table. This procedure was followed for all 15 unique pairs of the six competencies to produce the normative correlation matrix as reflected in Table 2.

Table 2

*Normative correlation matrix*

Normative Ratings	1	2	3	4	5	6
1. Leading Courageously	1	0.0716	0.1941	0.2639	0.4774	0.0118
2. Developing Self and Others	0.0716	1	0.48	0.0599	0.4204	0.2582
3. Delivering in our Global Environment	0.1941	0.48	1	0.1156	0.3221	0.3025
4. Drive Excellence and Innovation	0.2639	0.0599	0.1156	1	0.2306	0.2315
5. Communicate Honestly and Effectively	0.4774	0.4204	0.3221	0.2306	1	0.2506
6. Generate Customer Value	0.0118	0.2582	0.3025	0.2315	0.2506	1

**Correlation matrix of archival data.** Data were entered into SPSS 17.0 to obtain descriptive statistics and graphical plots to gain additional information on the quality and characteristics of the data. Table 3, the actual data matrix, shows the results of the further analysis of the archival data.

Table 3

*Actual Data Matrix*

Self-assessment Ratings	1	2	3	4	5	6
1. Leading Courageously	1	.750	.753	.817	.759	.686
2. Developing Self and Others	.750	1	.684	.683	.732	.548
3. Delivering in our Global Environment	.753	.684	1	.719	.692	.674
4. Drive Excellence and Innovation	.817	.683	.719	1	.666	.626
5. Communicate Honestly and Effectively	.759	.732	.692	.666	1	.562
6. Generate Customer Value	.686	.548	.674	.626	.562	1
Operational Manager Ratings	1	2	3	4	5	6
1. Leading Courageously	1	.842	.821	.849	.828	.705
2. Developing Self and Others	.842	1	.797	.754	.825	.611
3. Delivering in our Global Environment	.821	.797	1	.748	.776	.708
4. Drive Excellence and Innovation	.849	.754	.748	1	.685	.683
5. Communicate Honestly and Effectively	.828	.825	.776	.685	1	.606
6. Generate Customer Value	.705	.611	.708	.683	.606	1
Peer/Colleague Ratings	1	2	3	4	5	6
1. Leading Courageously	1	.828	.833	.837	.860	.728
2. Developing Self and Others	.828	1	.786	.713	.845	.609
3. Delivering in our Global Environment	.833	.786	1	.778	.813	.734
4. Drive Excellence and Innovation	.837	.713	.778	1	.712	.725
5. Communicate Honestly and Effectively	.860	.845	.813	.712	1	.649
6. Generate Customer Value	.728	.609	.734	.725	.649	1



Table 3 (continued)

Direct Report Ratings	1	2	3	4	5	6
1. Leading Courageously	1	.857	.860	.855	.852	.749
2. Developing Self and Others	.857	1	.811	.773	.860	.652
3. Delivering in our Global Environment	.860	.811	1	.808	.802	.787
4. Drive Excellence and Innovation	.855	.773	.808	1	.740	.733
5. Communicate Honestly and Effectively	.852	.860	.802	.740	1	.661
6. Generate Customer Value	.749	.652	.787	.733	.661	1

Customer/Client Ratings	1	2	3	4	5	6
1. Leading Courageously	1	.827	.842	.861	.852	.757
2. Developing Self and Others	.827	1	.792	.765	.836	.650
3. Delivering in our Global Environment	.842	.792	1	.800	.818	.776
4. Drive Excellence and Innovation	.861	.765	.800	1	.763	.741
5. Communicate Honestly and Effectively	.852	.836	.818	.763	1	.704
6. Generate Customer Value	.757	.650	.776	.741	.704	1

### Summary of Results

This section reports the results of the tests of the hypotheses associated with the study's six research questions. The rating data used in this study, described in Chapter 3, did not include any demographic or organizational information about either the raters or ratees. All that can be reported about these data are the numbers

of and means of the rates for each competency dimension by rater type. These values are reported in Table 4.

Table 4

*Numbers and Descriptive Statistics of Rates per Competency Dimension by Rater Type*

Competency Dimension	Respondent Type	Count	Mean	Standard Deviation
Lead Courageously	Self-assessment	1134	3.99	.39
	Ops. Manager	1008	4.05	.48
	Peers/Colleagues	4916	3.98	.53
	Direct Reports	4779	4.01	.55
	Customer/Client	3759	4.02	.52
Develop Self and Others	Self-assessment	1134	3.82	.43
	Ops. Manager	1008	3.86	.50
	Peers/Colleagues	4916	3.78	.57
	Direct Reports	4779	3.78	.63
	Customer/Client	3759	3.86	.55
Deliver in our Global Environment	Self-assessment	1134	3.85	.43
	Ops. Manager	1008	3.95	.48
	Peers/Colleagues	4916	3.97	.52
	Direct Reports	4779	4.03	.53
	Customer/Client	3759	4.02	.52
Drive Excellence and Innovation	Self-assessment	1134	3.77	.46
	Ops. Manager	1008	3.90	.51
	Peers/Colleagues	4916	3.87	.54
	Direct Reports	4779	3.87	.57
	Customer/Client	3759	3.89	.54
Communicate Honestly and Effectively	Self-assessment	1134	3.98	.41
	Ops. Manager	1008	4.07	.50
	Peers/Colleagues	4916	4.02	.55
	Direct Reports	4779	4.01	.58
	Customer/Client	3759	4.10	.53
Generate Customer Value	Self-assessment	1134	3.98	.51
	Ops. Manager	1008	4.12	.55
	Peers/Colleagues	4916	4.07	.57
	Direct Reports	4779	4.11	.57
	Customer/Client	3759	4.04	.60

Similarly, the questionnaire of the perceived co-occurrence rates of the six competency dimensions was administered to 32 managers and executives of a large firm under the guarantee of anonymity. No demographic or organizational information was gathered. These data are reported in terms of the perceived correlations derived. The researcher analyzed the data and examined the hypotheses in the present study by seeking answers to the following research questions:

### **Research Questions and Procedures**

1. Are the differences between the sources' correlation matrices for the actual data statistically significant?
2. Are the differences between the sources' correlation matrices for the actual data source types vs. the correlations for the normative group statistically significant?
3. What are the relative magnitudes of the differences between the correlation matrices from different sources?
4. How much difference do implicit performance theories make in the total rating scores that ratees receive?
5. How do the structures of the matrices for various sources differ?
6. How similar is the rotated factor structure (of the maximum number of non-error principal components) to the similarly rotated factor structure of the normative group?

The procedures for conducting the analyses for the foregoing research questions are detailed in Appendices E-J.

## Research Question 1 and Research Hypotheses

Are the differences between the sources' correlation matrices for the actual data statistically significant?

H<sub>1</sub>: There is significant difference between the correlation matrices for self vs. manager.

H<sub>2</sub>: There is significant difference between the correlation matrices for self vs. direct report.

H<sub>3</sub>: There is significant difference between the correlation matrices for self vs. peers.

H<sub>4</sub>: There is significant difference between the correlation matrices for self vs. customer/clients.

H<sub>5</sub>: There is significant difference between the correlation matrices for direct reports vs. peers.

H<sub>6</sub>: There is significant difference between the correlation matrices for peers vs. customer/clients.

H<sub>7</sub>: There is significant difference between the correlation matrices for manager vs. customers.

H<sub>8</sub>: There is significant difference between the correlation matrices for direct reports vs. peers.

H<sub>9</sub>: There is significant difference between the correlation matrices for direct reports vs. customer/clients.

H<sub>10</sub>: There is significant difference between the correlation matrices for peers vs. customer/clients.

The first research question asked whether the differences between the sources' correlation matrices were statistically significant. The pairs of matrices for each

combination of the five source types, self, subordinate, peer, manager, and customer/client, were compared using the asymptotic  $\chi^2$  test proposed by Jennrich (1970).

Table 5 presents the results of those tests and indices.

Table 5

*Results of Comparisons of Competency Dimension Correlation Matrices between Source Types*

Source Type	Compared to:	Jennrich $\chi^2$	<i>p</i>
Self	Manager	65.684	<.0001
Self	Peer	136.432	<.0001
Self	Direct Report	127.042	<.0001
Self	Customer/Client	87.506	<.0001
Manager	Peer	60.328	<.0001
Manager	Direct Report	45.219	<.0001
Manager	Customer/Client	73.336	<.0001
Peer	Direct Report	60.888	<.0001
Peer	Customer/Client	41.483	0.0003
Direct Report	Customer/Client	44.178	0.0001

*Note:*  $df = 15$  for all Jennrich  $\chi^2$  tests.

The results in Table 5 indicate that the difference between every pair of matrices is statistically significant. Research hypotheses 1-10 examined whether differences between the sources' correlation matrices for the actual data are statistically significant. There is sufficient evidence to support that there are significant correlations between the matrices for each pair of source types. Note (per Table 5), however, that the differences are not large in an absolute sense. The absolute magnitudes of these differences were addressed in reference to research question 3.

## Research Question 2 and Research Hypotheses

Are the differences between the sources' correlation matrices for the actual data source types vs. the correlations for the normative group statistically significant?

H<sub>11</sub>: There is significant difference between the correlation matrix for self vs. the correlations for the normative group.

H<sub>12</sub>: There is significant difference between the correlation matrix for direct reports vs. the correlations for the normative group.

H<sub>13</sub>: There is significant difference between the correlation matrix for managers vs. the correlations for the normative group.

H<sub>14</sub>: There is significant difference between the correlation matrix for peers vs. the correlations for the normative group.

H<sub>15</sub>: There is significant difference between the correlation matrix for customer/clients vs. the correlations for the normative group.

The second research question asked whether the differences between the normative correlation matrix, that is, derived from the questionnaire of perceived covariation, and the correlation matrices of the five source types are statistically significant. The normative correlation matrix was compared to each source type matrix using the asymptotic  $\chi^2$  test proposed by Jennrich (1970). Table 6 presents the results of these tests and indices.

Table 6

*Results of Comparisons of the Normative and Source Type Competency Dimension Correlation Matrices*

Normative Compared to:	Jennrich $\chi^2$	P
Self	832.324	<.0001
Manager	511.2374	<.0001
Peer	881.638	<.0001
Direct Report	902.360	<.0001
Customer/Client	641.240	<.0001

*Note: df = 15 for all Jennrich  $\chi^2$  tests.*

The differences between the source type and normative correlation matrices were both significant and large in an absolute sense. Research hypotheses 11-15 examine whether differences between the normative correlation matrix and the correlation matrices of the five source types were statistically significant. There is sufficient evidence to support that there are significant differences between the normative correlation matrix and the correlation matrices of the five source types.

**Research Question 3 and Sub-Research Questions**

What are the relative magnitudes of the differences between the correlation matrices from various sources?

- 3a. What is the relative magnitude of difference between the correlation matrices for self vs. manager?
- 3b. What is the relative magnitude of difference between the correlation matrices for self vs. direct reports?
- 3c. What is the relative magnitude of difference between the correlation matrices for self vs. peers?

- 3d. What is the relative magnitude of difference between correlation matrices for self vs. customer/clients?
- 3e. What is the relative magnitude of difference between the correlation matrices for manager vs. direct reports?
- 3f. What is the relative magnitude of difference between the correlation matrices for manager vs. peers?
- 3g. What is the relative magnitude of difference between the correlation matrices for manager vs. customer/clients?
- 3h. What is the relative magnitude of difference between correlation matrices for direct reports vs. peers?
- 3i. What is the relative magnitude of difference between the correlation matrices for direct reports vs. customer/clients?
- 3j. What is the relative magnitude of difference between the correlation matrices for peers vs. customer/clients?

The third research question sought to elucidate the absolute magnitude of the differences between the correlation matrices, both among the source types and between the normative and source type groups. This was accomplished by computing the Euclidean distances of the correlations in the lower triangular portions of the matrices being compared. The results of these computations are presented in Tables 7 and 8.



Table 7

*Euclidean Distances Between the Lower Triangular Portions of the Source Type**Correlation Matrices*

Source Type	Compared to:	Euclidean Distance
Self	Manager	0.253
Self	Peer	0.306
Self	Direct Report	0.385
Self	Customer/Client	0.380
Manager	Peer	0.107
Manager	Direct Report	0.164
Manager	Customer/Client	0.184
Peer	Direct Report	0.116
Peer	Customer/Client	0.119
Direct Report	Customer/Client	0.072

An average difference of the 10 pairs of corresponding correlations would produce the largest of the Euclidean distances among the 10-pair comparisons between Self and Direct Report,  $D = 0.385$ , reported in Table 7. In the case of the smallest Euclidean distance, that is, between Direct Report and Customer/Client,  $D = 0.072$ , the average difference between corresponding correlations was only 0.023. Overall, the Euclidean distances indicate that the average difference among corresponding correlations is on the order of 0.066.

Table 8

*Euclidean Distances Between the Lower Triangular Portions of the Normative and*

*Source Type Correlation Matrices*

Normative Compared to:	Euclidean Distance
Self	1.818
Manager	2.027
Peer	2.072
Direct Report	2.164
Customer/Client	2.156

Table 8 shows the mean Euclidean distance between the normative matrix and the five source type matrices to be 2.047, which translates to an average difference between pairs of correlations of 0.529. The reason for this large difference was that the normative group saw the six competency dimensions as being much more independent of each other than did the various source types. The average correlation between dimensions in the normative matrix was 0.246, whereas the average inter-dimensional correlation among the source types was 0.762. It can be concluded that the difference between the normative and source type inter-dimensional correlations is consistently large across all source types.

#### **Research Question 4**

How much difference do different implicit performance theories make in the total rating scores that ratees receive?

### Research Question 4A and Research Hypotheses

4a. Are there differences in the correlations between weighted mean scores for source types and the predicted scores based on source types' regression equation?

H<sub>16</sub>: There is a difference in the correlations between weighted mean scores for source type self and the predicted scores based on source type-managers' regression equation.

H<sub>17</sub>: There is a difference in the correlations between weighted mean scores for source type-self and the predicted scores based on source type-direct reports' regression equation.

H<sub>18</sub>: There is a difference in the correlations between weighted mean scores for source type-self and the predicted scores based on source type-peers' regression equation.

H<sub>19</sub>: There is a difference in the correlations between weighted mean scores for source type-self and the predicted scores based on source type-customer/clients' regression equation.

H<sub>20</sub>: There is a difference in the correlations between weighted mean scores for source type- manager and the predicted scores based on source type-direct reports' regression equation.

H<sub>21</sub>: There is a difference in the correlations between weighted mean scores for source type- manager and the predicted scores based on source type-peers' regression equation.

H<sub>22</sub>: There is a difference in the correlations between weighted mean scores for source type- manager and the predicted scores based on source type-customer/clients' regression equation.

H<sub>23</sub>: There is a difference in the correlations between weighted mean scores for source type-direct reports' and the predicted scores based on source type-peers' regression equation.

H<sub>24</sub>: There is a difference in the correlations between weighted mean scores for source type-direct reports' and the predicted scores based on source type-customer/clients' regression equation.

H<sub>25</sub>: There is a difference in the correlations between weighted mean scores for source type-peer and the predicted scores based on source type-customer/clients' regression equation.

Research question 4 sought to determine whether, for each source type, there are differences in the correlations between, and the means of, actual total scores and the predicted scores based on the other source types' regression equations. This question explored the issue of whether differences between the rating policies of the different source types make a difference in the overall total scores across all the competency dimensions. In actuality, the weighted mean score rather than the total score was used in order to ensure that the overall score is on the same scale as the individual dimension scores. In addition, the standardized regression weights and standard scores on the competency dimensions were used in order to eliminate the effect of mean and variance differences between source types. The resulting differences are consequently solely due

to the effect of different rating source policies. A *rating source policy* refers to a judgment policy and is the relative weight assigned to the competencies by a source type.

The results of the correlational analyses were identical. The correlations between the mean scores (across the six competency dimensions) for each source type and the weighted means using the regression weights of the other source types rounded to 1.0 in every case. The failure to observe any differences in these analyses was due to the near-equality of the means and standard deviations of the six competency dimensions within each source type. The homogeneity of the standard deviations resulted in very small differences between the regression weights. The application of these small weight differences to competency dimension scores which differed very little in their means resulted in there being virtually no effect on the ordering of the weighted means. Research hypotheses 16-25 explored the issue of whether differences between the rating policies of the different source types make a difference in the weighted mean score across all the competency dimensions. Thus, it can be concluded that there is insufficient evidence to support that differences between the rating policies of the various source types make a difference in the weighted mean score across all the competency dimensions.

The tests of the differences in means were conducted using the paired t-test. The results of these tests are reported in Table 9.

#### **Research Question 4B and Research Hypotheses**

- 4b. Are there differences in the means between actual total scores for source types and the predicted scores based on source types' regression equation?

- H<sub>26</sub>: There is a difference in the means between actual total scores for source type-self and the predicted scores based on source type-managers' regression equation.
- H<sub>27</sub>: There is a difference in the means between actual total scores for source type-self and the predicted scores based on source type-direct reports' regression equation.
- H<sub>28</sub>: There is a difference in the means between actual total scores for source type-self and the predicted scores based on source type-peers' regression equation.
- H<sub>29</sub>: There is a difference in the means between actual total scores for source type self and the predicted scores based on source type-customer/clients' regression equation.
- H<sub>30</sub>: There is a difference in the means between actual total scores for source type-manager and the predicted scores based on source type-direct reports' regression equation.
- H<sub>31</sub>: There is a difference in the means between actual total scores for source type-manager and the predicted scores based on source type-peers' regression equation.
- H<sub>32</sub>: There is a difference in the means between actual total scores for source type-manager and the predicted scores based on source type-customer/clients' regression equation.
- H<sub>33</sub>: There is a difference in the means between actual total scores for source type-direct reports' and the predicted scores based on source type-peers' regression equation.
- H<sub>34</sub>: There is a difference in the means between actual total scores for source type-direct reports and the predicted scores based on source type-customer/clients' regression equation.

H<sub>35</sub>: There is a difference in the means between actual total scores for source type-peer and the predicted scores based on source type-customer/clients' regression equation.

Table 9

*t-test Results for Mean Differences Between Each Source Type's Actual Overall Mean Scores and Those Formed With the Weights of Each Other Source Type*

Mean of Overall Mean Scores for:	Minus Mean of Weighted Mean Scores Based on Weights of:	Differences in Means			df	Sig. (2-tailed)
		Mean	Std. Error	t		
Self	Manager	-.023	.0013	-18.418	825	<.001
	Peers	-.024	.0013	-18.416	825	<.001
	Direct Reports	-.024	.0019	-12.440	825	<.001
	Customer/Client	-.023	.0017	-13.406	825	<.001
Manager	Self	.011	.0023	4.677	489	<.001
	Peers	.011	.0008	14.337	489	<.001
	Direct Reports	.010	.0005	18.407	489	<.001
	Customer/Client	.010	.0003	30.887	489	<.001
Peers	Self	-.008	.0007	-11.163	912	<.001
	Manager	-.007	.0011	-6.779	912	<.001
	Direct Reports	-.007	.0015	-4.843	912	<.001
	Customer/Client	-.007	.0016	-4.673	912	<.001
Direct Reports	Self	-.016	.0016	-9.758	851	<.001
	Manager	-.015	.0006	-26.471	851	<.001
	Peers	-.015	.0003	-45.236	851	<.001
	Customer/Client	-.015	.0007	-20.675	851	<.001
Customer/Client	Self	.001	.0017	.789	656	.430
	Manager	.002	.0003	7.398	656	<.001
	Peers	.002	.0005	4.728	656	<.001
	Direct Reports	.003	.0007	2.984	656	.003

The results in Table 9 indicated that there is sufficient evidence to support research hypotheses 26-33 and 35, that is, that there are significant differences in means. Only in the case of self weightings applied to customer/clients with self and direct reports (research hypothesis 34), there was no significant difference in the means of the resulting

scores. However, these differences, while statistically significant due to the large sample size, are minuscule. Given that the standard deviations of the variables being compared were approximately 1.0 in all cases, the effect is less than 2.5% in all cases.

### Research Question 5

How do the structures of the matrices for various sources differ?

The fifth research question inquired as to the comparability between source types of the composition of the optimum number of rotated principal components computed for each source type's ratings of the six competency dimensions. These comparisons consisted of pairing the rotated factors that were most similar between source types in terms of the competency dimensions that defined them and then computing the correlation between the rotated factor loadings for each pair of factors. Results of these analyses are reported in Table 10. Note that shading indicates dominant variables within the factor.

Table 10

*Comparisons between Pairs of Most Similar Rotated Factors between Source Types*

Self vs. Mgr.	Self_1	Mgr_1	Self_2	Mgr_3	Self_3	Mgr_2	Self_4	Mgr_4	Self_5	Mgr_5
1. Lead Courageously	0.652	0.598	0.345	0.306	0.436	0.484	0.313	0.395	0.256	0.287
2. Develop Self and Others	0.325	0.361	0.206	0.231	0.329	0.392	0.826	0.762	0.247	0.283
3. Deliver in our Global Environment	0.352	0.355	0.315	0.359	0.277	0.36	0.27	0.33	0.792	0.712
4. Drive Excellence and Innovation	0.83	0.823	0.259	0.334	0.224	0.243	0.268	0.27	0.283	0.243
5. Comm. Honestly and Effectively	0.299	0.289	0.233	0.28	0.825	0.801	0.321	0.336	0.248	0.272
6. Generate Customer Value	0.278	0.304	0.887	0.873	0.21	0.241	0.18	0.189	0.241	0.227
Correlations	0.993		0.988		0.991		0.984		0.988	



Table 10 (continued)

Self vs. Peer	Self_1	Peer_1	Self_2	Peer_3	Self_3	Peer_2	Self_4	Peer_4	Self_5	Peer_5
1. Lead Courageously	0.652	0.527	0.345	0.359	0.436	0.551	0.313	0.379	0.256	0.285
2. Develop Self and Others	0.325	0.297	0.206	0.226	0.329	0.394	0.826	0.797	0.247	0.265
3. Deliver in our Global Environment	0.352	0.348	0.315	0.339	0.277	0.362	0.27	0.324	0.792	0.726
4. Drive Excellence and Innovation	0.83	0.82	0.259	0.342	0.224	0.251	0.268	0.262	0.283	0.258
5. Comm. Honestly and Effectively	0.299	0.263	0.233	0.26	0.825	0.772	0.321	0.401	0.248	0.297
6. Generate Customer Value	0.278	0.306	0.887	0.872	0.21	0.232	0.18	0.197	0.241	0.228
Correlations	0.975		0.995		0.971		0.985		0.990	

Self vs. Direct Reports	Self_1	DR_3	Self_2	DR_2	Self_3	DR_1	Self_4	Self_5
1. Lead Courageously	0.652	0.525	0.345	0.43	0.436	0.677	0.313	0.256
2. Develop Self and Others	0.325	0.379	0.206	0.29	0.329	0.828	0.826	0.247
3. Deliver in our Global Environment	0.352	0.392	0.315	0.584	0.277	0.621	0.27	0.792
4. Drive Excellence and Innovation	0.83	0.804	0.259	0.391	0.224	0.433	0.268	0.283
5. Comm. Honestly and Effectively	0.299	0.28	0.233	0.322	0.825	0.866	0.321	0.248
6. Generate Customer Value	0.278	0.305	0.887	0.886	0.21	0.316	0.18	0.241
Correlations	0.964		0.940		0.730			

Table 10 (continued)

Self vs. Cust/ Client	Self 1	Cust/ Client 3	Self 2	Cust/ Client 2	Self 3	Cust/ Client 1	Self 4	Self 5
1. Lead Courageously	0.652	0.559	0.345	0.451	0.436	0.631	0.313	0.256
2. Develop Self and Others	0.325	0.388	0.206	0.243	0.329	0.838	0.826	0.247
3. Deliver in our Global Environment	0.352	0.384	0.315	0.585	0.277	0.617	0.27	0.792
4. Drive Excellence and Innovation	0.83	0.808	0.259	0.398	0.224	0.414	0.268	0.283
5. Comm. Honestly and Effectively	0.299	0.279	0.233	0.406	0.825	0.819	0.321	0.248
6. Generate Customer Value	0.278	0.334	0.887	0.869	0.21	0.322	0.18	0.241
Correlations	0.974		0.922		0.669			

Mgr vs. Peer	Mgr 1	Peer 1	Mgr 2	Peer 2	Mgr 3	Peer 3	Mgr 4	Peer 4	Mgr 5	Peer 5
1. Lead Courageously	0.598	0.527	0.484	0.551	0.306	0.359	0.395	0.379	0.287	0.285
2. Develop Self and Others	0.361	0.297	0.392	0.394	0.231	0.226	0.762	0.797	0.283	0.265
3. Deliver in our Global Environment	0.355	0.348	0.36	0.362	0.359	0.339	0.33	0.324	0.712	0.726
4. Drive Excellence and Innovation	0.823	0.82	0.243	0.251	0.334	0.342	0.27	0.262	0.243	0.258
5. Comm. Honestly and Effectively	0.289	0.263	0.801	0.772	0.28	0.26	0.336	0.401	0.272	0.297
6. Generate Cust Value	0.304	0.306	0.241	0.232	0.873	0.872	0.189	0.197	0.227	0.228
Correlations	0.989		0.988		0.993		0.990		0.997	

Table 10 (continued)

Mgr. vs. Dir. Reports	Mgr 1	DR 3	Mgr 2	DR 1	Mgr 3	DR 2	Mgr 4	Mgr 5
1. Lead Courageously	0.598	0.525	0.484	0.677	0.306	0.43	0.395	0.287
2. Develop Self and Others	0.361	0.379	0.392	0.828	0.231	0.29	0.762	0.283
3. Deliver in our Global Environment	0.355	0.392	0.36	0.621	0.359	0.584	0.33	0.712
4. Drive Excellence and Innovation	0.823	0.804	0.243	0.433	0.334	0.391	0.27	0.243
5. Comm. Honestly and Effectively	0.289	0.28	0.801	0.866	0.28	0.322	0.336	0.272
6. Generate Customer Value	0.304	0.305	0.241	0.316	0.873	0.886	0.189	0.227
Correlations	0.986		0.794		0.946			

Manager vs. Customer/Client	Mgr 1	Cust/ Client 3	Mgr 2	Cust/ Client 1	Mgr 3	Cust/ Client 2	Mgr 4	Mgr 5
1. Lead Courageously	0.598	0.559	0.484	0.631	0.306	0.451	0.395	0.287
2. Develop Self and Others	0.361	0.388	0.392	0.838	0.231	0.243	0.762	0.283
3. Deliver in our Global Environment	0.355	0.384	0.36	0.617	0.359	0.585	0.33	0.712
4. Drive Excellence and Innovation	0.823	0.808	0.243	0.414	0.334	0.398	0.27	0.243
5. Comm. Honestly and Effectively	0.289	0.279	0.801	0.819	0.28	0.406	0.336	0.272
6. Generate Customer Value	0.304	0.334	0.241	0.322	0.873	0.869	0.189	0.227
Correlations	0.993		0.738		0.93			

Table 10 (continued)

Peer vs. Dir. Rep'ts (DR)	Peer_1	DR_3	Peer_2	DR_1	Peer_3	DR_2	Peer_4	Peer_5
1. Lead Courageously	0.527	0.525	0.551	0.677	0.359	0.43	0.379	0.285
2. Develop Self and Others	0.297	0.379	0.394	0.828	0.226	0.29	0.797	0.265
3. Deliver in our Global Environment	0.348	0.392	0.362	0.621	0.339	0.584	0.324	0.726
4. Drive Excellence and Innovation	0.82	0.804	0.251	0.433	0.342	0.391	0.262	0.258
5. Comm. Honestly and Effectively	0.263	0.28	0.772	0.866	0.26	0.322	0.401	0.297
6. Generate Customer Value	0.306	0.305	0.232	0.316	0.872	0.886	0.197	0.228
Correlations	0.989		0.801		0.939			

Peer vs. Cust/Client	Peer 1	Cust/ Client 3	Peer 2	Peer 3	Cust/ Client 2	Peer 4	Cust/ Client 1	Peer 5
1. Lead Courageously	0.527	0.559	0.551	0.359	0.451	0.379	0.631	0.285
2. Develop Self and Others	0.297	0.388	0.394	0.226	0.243	0.797	0.838	0.265
3. Deliver in our Global Environment	0.348	0.384	0.362	0.339	0.585	0.324	0.617	0.726
4. Drive Excellence and Innovation	0.82	0.808	0.251	0.342	0.398	0.262	0.414	0.258
5. Comm. Honestly and Effectively	0.263	0.279	0.772	0.26	0.406	0.401	0.819	0.297
6. Generate Customer Value	0.306	0.334	0.232	0.872	0.869	0.197	0.322	0.228
Correlations	0.991			0.921		0.794		

Table 10 (continued)

Dir. Repts vs. Customer/Client	Cust/		Cust/		Cust/	
	DirRpt_1	Client_1	DirRpt_2	Client_2	DirRpt_3	Client_3
1. Lead Courageously	0.677	0.631	0.43	0.451	0.525	0.559
2. Develop Self and Others	0.828	0.838	0.29	0.243	0.379	0.388
3. Deliver in our Global Environment	0.621	0.617	0.584	0.585	0.392	0.384
4. Drive Excellence and Innovation	0.433	0.414	0.391	0.398	0.804	0.808
5. Comm. Honestly and Effectively	0.866	0.819	0.322	0.406	0.28	0.279
6. Generate Customer Value	0.316	0.322	0.886	0.869	0.305	0.334
Correlations	0.994		0.980		0.996	

The comparisons in Table 10 can be summarized as follows: The factor structures of the self, manager, and peer ratings consisted of five factors that were identical across these three source types. The ratings of all three source types had factors dominated by the same variables. All three source types divided the Lead Courageously dimension between two factors dominated by the Drive Excellence and Innovation dimension and by the Communicate Honestly and Effectively dimension, respectively. Three of the five factors were dominated by a single competency dimension, and the other two were dominated by two competency dimensions, one of which, in both cases, was Lead Courageously. The mean correlation between paired factors across all three groups was 0.988. The factor structures of the direct reports and customer/clients consisted of only three factors, dominated by the same variables for both source types. The Lead Courageously dimension was divided across all three factors, and Deliver in

our Global Environment was divided across two of the factors. One of the three factors contained the dominant loadings for four of the six competency dimensions. A second factor was dominated by three competency dimensions, two of which (Lead Courageously and Deliver in Our Global Environment) were shared with the aforementioned factor. The third factor was dominated by two competency dimensions, one of which was Lead Courageously. The mean correlation between paired factors across these two source type groups was 0.990.

### **Research Question 6**

How similar is the rotated factor structure (of the maximum number of non-error principal components) to the similarly rotated factor structure of the normative group?

The sixth and final research question inquired as to the comparability between the optimum rotated factor structure derived from the normative correlations among the six competency dimensions and the optimum rotated factor structure of competency dimension ratings of each of the five source types. Once again, these comparisons consisted of pairing up the rotated factors that were most similar between the normative factor solution and that of each source type in terms of the competency dimensions that defined them, and then computing the correlation between the rotated factor loadings for each pair of factors. The results of these analyses are reported in Table 11. Note that shading indicates dominant variables within the factor.

Table 11

*Comparisons between Pairs of Most Similar Rotated Factors for the Normative Factor*

*Solution and that for Each Source Type*

Normative vs. Self	Norm 1	Self 4	Norm 2	Self 3	Norm 3	Self 2	Norm 4	Self 1	Self 5
1. Lead Courageously	0.008	0.313	0.882	0.436	-0.126	0.345	0.206	0.652	0.256
2. Develop Self and Others	0.845	0.826	0.129	0.329	0.148	0.206	-0.079	0.325	0.247
3. Deliver in our Global Environment	0.836	0.27	0.109	0.277	0.086	0.315	0.138	0.352	0.792
4. Drive Excellence and Innovation	0.042	0.268	0.157	0.224	0.128	0.259	0.961	0.83	0.283
5. Comm. Honestly and Effectively	0.335	0.321	0.783	0.825	0.259	0.233	0.001	0.299	0.248
6. Generate Customer Value	0.176	0.18	0.034	0.21	0.957	0.887	0.131	0.278	0.241
Correlations	0.596		0.783		0.865		0.868		

Normative vs. Manager	Norm 1	Mgr 4	Norm 2	Mgr 2	Norm 3	Mgr 3	Norm 4	Mgr 1	Mgr 5
1. Lead Courageously	0.008	0.395	0.882	0.484	-0.126	0.306	0.206	0.598	0.287
2. Develop Self and Others	0.845	0.762	0.129	0.392	0.148	0.231	-0.079	0.361	0.283
3. Deliver in our Global Environment	0.836	0.33	0.109	0.36	0.086	0.359	0.138	0.355	0.712
4. Drive Excellence and Innovation	0.042	0.27	0.157	0.243	0.128	0.334	0.961	0.823	0.243
5. Comm. Honestly and Effectively	0.335	0.336	0.783	0.801	0.259	0.28	0.001	0.289	0.272
6. Generate Customer Value	0.176	0.189	0.034	0.241	0.957	0.873	0.131	0.304	0.227
Correlations	0.6		0.788		0.907		0.898		

Table 11 (continued)

Normative vs. Peers	Norm 1	Peer 4	Norm 2	Peer 2	Norm 3	Peer 3	Norm 4	Peer 1	Peer 5
1. Lead Courageously	0.008	0.379	0.882	0.551	-0.126	0.359	0.206	0.527	0.285
2. Develop Self and Others	0.845	0.797	0.129	0.394	0.148	0.226	-0.079	0.297	0.265
3. Deliver in our Global Environment	0.836	0.324	0.109	0.362	0.086	0.339	0.138	0.348	0.726
4. Drive Excellence and Innovation	0.042	0.262	0.157	0.251	0.128	0.342	0.961	0.82	0.258
5. Comm. Honestly and Effectively	0.335	0.401	0.783	0.772	0.259	0.26	0.001	0.263	0.297
6. Generate Customer Value	0.176	0.197	0.034	0.232	0.957	0.872	0.131	0.306	0.228
Correlations	0.612		0.864		0.866		0.953		

Normative vs. Dir. Rpts	Norm 1	Norm 2	Dir Rpt 1	Norm 3	Dir Rpt 2	Norm 4	Dir Rpt 3
1. Lead Courageously	0.008	0.882	0.677	-0.126	0.43	0.206	0.525
2. Develop Self and Others	0.845	0.129	0.828	0.148	0.29	-0.079	0.379
3. Deliver in our Global Environment	0.836	0.109	0.621	0.086	0.584	0.138	0.392
4. Drive Excellence and Innovation	0.042	0.157	0.433	0.128	0.391	0.961	0.804
5. Comm. Honestly and Effectively	0.335	0.783	0.866	0.259	0.322	0.001	0.28
6. Generate Customer Value	0.176	0.034	0.316	0.957	0.886	0.131	0.305
Correlations		0.544		0.768		0.927	



Table 11 (continued)

Normative vs. Customer/ Client	Norm 1	Cust/ Client 1	Norm 2	Norm 3	Cust/ Client 2	Norm 4	Cust/ Client 3
1. Lead Courageously	0.008	0.631	0.882	-0.126	0.451	0.206	0.559
2. Develop Self and Others	0.845	0.838	0.129	0.148	0.243	-0.079	0.388
3. Deliver in our Global Environment	0.836	0.617	0.109	0.086	0.585	0.138	0.384
4. Drive Excellence and Innovation	0.042	0.414	0.157	0.128	0.398	0.961	0.808
5. Comm. Honestly and Effectively	0.335	0.819	0.783	0.259	0.406	0.001	0.279
6. Generate Customer Value	0.176	0.322	0.034	0.957	0.869	0.131	0.334
Correlations		0.548			0.757		0.917

The normative factor structure differed both from the five-factor structure of the self, manager, and peer raters, and from the three-factor structure of the direct report and customer/client raters. It consisted of four factors. The normative group saw Deliver in our Global Environment to be part of the same underlying construct as Develop Self and Others, whereas the self, manager, and peer raters saw it as representing a completely separate construct. The normative group also saw Lead Courageously to be strongly linked with Communicate Honestly and Effectively in defining a common underlying construct. In contrast, the self, manager, and peer raters saw Lead Courageously to be almost evenly split between the Communicate factor and the Drive Excellence and Innovation factor. The only factor on which the normative group and the self, manager, and peer groups fully agreed as to its composition was the one defined by Generate

Customer Value. The average correlation between the normative factor loadings and those of the latter three rater groups was 0.80.

The ratings by the other two rating source type groups, direct reports and customer/client, produced factors which, although highly similar between these two source type groups, departed quite substantially from those of the normative group. The one additional factor that did not emerge in either the direct report or customer/client source type group was different for each group. The direct reports did not see the factor that the normative group defined in terms of the Develop Self and Others and Deliver in our Global Environment dimensions. Instead, they included these dimensions in a highly general factor that also included Lead Courageously, Drive Excellence and Innovation, and Communicate Honestly and Effectively.

The customer/client source group included the Develop Self and Others and Deliver in our Global Environment dimensions on the same factor as did the normative group, but also diverged from the normative group by including in the same factor the Lead Courageously and Communicate Honestly and Effectively dimensions. The customer/client source group did not perceive the factor that the normative group defined exclusively in terms of Lead Courageously and Communicate Honestly and Effectively. Also, whereas the normative group defined one factor exclusively in terms of the Generate Customer Value dimension, the customer/client source type group saw this factor as a much more complex combination of dimensions that additionally included the Deliver in our Global Environment dimension and lesser amounts of two other dimensions.

Both direct reports and customer/clients defined the structure of the factor dominated by Drive Excellence and Innovation most similarly to the normative group, but the former two groups also perceived Lead Courageously to a noticeable influence in the underlying construct. The substantial disagreement between these two source type groups and the normative group is reflected in the mean correlation across all the factors of 0.744.

### **Conclusion**

This study's research question, "What is the nature of the differences between performance ratings from different rating sources?" examined whether multisource feedback results are true reflections of what raters think and how their opinions match with a normative pattern structure. A significant number of other studies have provided rationale for source differences with no definitive conclusions (Hoffman, 2006; Liff, 2010; Waldman; Atwater; & Antonioni, 1998). This chapter described the statistical analysis and findings of the present study, including details about the actual archival data sample.

An analysis of the statistics corresponding to the research hypotheses 1-15 revealed that the correlation matrices of all the source types and the normative structure differed significantly. The absolute magnitude of the differences between the correlation matrices, both among the source types and between the normative and source type groups, was measured. The results were presented in tables 7 and 8. Additionally, the raters' rating policy differences made no difference in the ordering of the scores: All the correlations were 1.0, meaning that differences between the raters were almost completely due to the differences in the way they actually rated people. The self,

manager, and peer structures are virtually identical. The direct report and customer/client structures are also virtually identical to each other, but differentiate two less factors than the first group. However, the normative group structure consisted of four factors. One factor that is virtually identical across all groups is the combination of Lead Courageously and Drive Excellence and Innovation. The normative group saw Deliver in our Global Environment to be part of the same underlying construct as Develop Self and Others and Lead Courageously to be strongly linked with Communicate Honestly and Effectively in defining a common underlying construct.

Chapter 5 expands on the findings reported in Chapter 4 to provide additional analysis of the results and includes discussions of what interpretations may account for the kind of differences and the potential for additional research to expand on the relationship between the normative pattern and archival pattern of ratings of multisource feedback assessments.

## CHAPTER 5: DISCUSSION, IMPLICATIONS, RECOMMENDATIONS

### Discussion

#### Summary of Study

For decades organizations have sought to assess the performance of their employees. The 360 degree feedback assessment process, wherein multiple raters complete an instrument to rate an individual's competencies, dates from the 1950 and 1960 human relations movements (Waldman, Atwater, & Antonioni, 1998). The process gained currency in the 1970s when Clark Wilson, an industrial psychologist, introduced the multisource feedback concept to management training. His research and development of early survey instruments transformed the way that businesses evaluated their managers. In 1997, Lepsinger and Lucia wrote that all Fortune 500 firms were either using or thinking about using 360 degree feedback.

However, rating accuracy is often in question. Some view performance appraisals as inherently biased because basic to the process are personal judgments, subjective principles, and individual viewpoints (Thiry, 2009, p. 3). Biases may bring rating error to performance assessments, causing rating inaccuracies. Fletcher, Baldry, and Cunningham-Snell (1998) concluded that different rater groups, affected by their own viewpoints, tend to make somewhat different assessments. Although many studies raised the question of rater biases and accuracy of multisource assessments, they left open the issue of rater differences and the reason for different ratings from different rater sources. This researcher found no studies that examined source differences by comparing a normative correlation matrix pattern of ratings with a correlation matrix of actual ratings in multisource feedback, leading to the following research question being explored for

this study: “What is the nature of the differences between performance ratings from different rating sources?”

This study used archival data and a constructed normative pattern to compare source differences and interpret the meaning of source factors to add to the body of knowledge on source differences. In exploring the relationship between a normative pattern and an archival pattern of findings, the study examined interrater agreement and disagreement, including the pattern, if any, of differences between rater sources. The research posited that the pattern of correlation on items of memory-based multisource feedback instruments might reveal more about each source type’s cognitive framework than about “what” correlates with “what” across individual differences in performance.

### **Summary of Findings and Interpretation of Results**

The research results are discussed in the context of two questions. Research questions 1-3 are discussed as a group under the following question: “Are there differences between the archival data and a normative pattern, and, if so, what are the differences, and how do they compare to previous research?” Research questions 4-6 are discussed under the following question: “What is the pattern, if any, of differences between rater sources and rater sources and the normative structure?” In answering those questions, the researcher emphasizes the implications and limitations of the study and opportunities for future research.

**Findings, Implications, and Limitations: Research Questions 1-3.** In answer to the first research question, “Are the differences between the sources’ correlation matrices for the actual data statistically significant?” research hypotheses 1-10 examine

whether differences between the sources' correlation matrices for the actual data are statistically significant and if there is sufficient evidence to support that there are significant correlations between the matrices for each pair of source types. The results indicate that the difference between every pair of source types within the correlation matrices is statistically significant.

Although there are various speculations about their meaning, the origins of source differences are little understood, according to Hoffman (2006). For example, Campbell and Lee (1998) proposed that differences in ratings across sources can be attributed to varying concepts of effective performance by diverse sources. Borman (1974) suggested that the differences resulted from different opportunities to observe the target behavior. Beauvois and Dubois (2000) posited that the differences stemmed from the benefits of the target's behavior to a given rater. Lance and Woehr (1989) wrote that the differences could result from the display of different behaviors in the presence of different groups of raters. Hoffman (2006) conjectured that source effects simply represent variance that cannot be attributed to a target's performance. A frequent explanation in other studies of source effects is that some raters have a greater opportunity to see certain behaviors than other raters.

The present study's findings regarding source differences align with prior research. The ecological hypothesis implies that managers are intentional in choosing which behaviors to display in the presence of different raters and according to the demands of different situations (Lance & Woehr, 1989; Salam, Cox, & Sims, 1997). One might expect that if someone greatly values a certain behavior, the display of that behavior would have a significant impact on one's general impression of a coworker

(Hoffman, 2006, p. 127). The results of this research question support previous studies regarding source differences but leave unanswered the questions, “Are raters seeing the same performance differently? Or do raters perceive different samples of performance by a given rater?”

The second research question asks, “Are the differences between the sources’ correlation matrices for the actual data source types vs. the correlations for the normative group statistically significant?” This question and hypotheses 11-15 explored whether the differences between the normative correlation matrix, that is, derived from a questionnaire of perceived covariation, and the correlation matrices of the five source types were statistically significant. The normative correlation matrix and a source type matrix were compared via the Jennrich test, and the findings indicated that differences between the source types and normative correlation matrices were statistically significant. There is sufficient evidence to support that there are significant differences between the normative correlation matrix and the correlation matrices of the five source types.

This research question examined whether multisource feedback results are true reflections of what the raters think and how their opinions matched with a normative pattern from a second group of raters. The findings of statistically significant differences between the normative correlation matrix and the correlation matrices of actual data support the view that executives who used the normative structure in their ratings may not have been influenced by a common conceptual organization of “what goes with what?” Instead of giving ratings that matched memory-based ratings, they may have based their ratings on their own unique conceptual schema. The results of this research question diverge from previous research in other fields that used a normative pattern to



compare with actual data. The present data do not reflect the same results of previous studies from researchers who concluded that, if conceptual associations strongly affect memory-based rating and if conceptual associations are not an accurate reflection of true covariance, then dimension analyses may reveal more about conceptual associations or beliefs about inter-item correlations than about trait behavior covariations (Lodzinski, 1991, p. 4).

Shweder & D'Andrade (1980) further stated that patterns of correlation among items on memory-based personality instruments reveal more about diverse forms of implicit person theory in the minds of raters than about “what” actually correlates with “what” across individual differences in conduct. They concluded that such forms of conceptual affiliation could be most easily discovered by simply asking a handful of informants, “What is like what?” (p. 38). Thirty minutes of videotaped interaction among members of a family revealed that memory-based rating structures paralleled pre-existing similarity-of-meaning structures but did not accurately reflect the correlation structure of actual behavior (Shweder & D'Andrade, 1980, p. 54). Shweder (1982) also posited that the inter-item correlational structure of memory-based rating may be more a reflection of raters' schematically organized inter-item associations than the true pattern and extent of covariance among the dimensions, traits, or behavior to which the items refer. Shweder reached that conclusion by examining the correlational structure of psychiatric ratings using categories from the Brief Psychiatric Rating Scale, which revealed that this rating structure could be reproduced from judgments about “similarity in meaning.”

To summarize, this research study adds to previous research by departing from previous studies that supported the view that a correlation matrix of memory-based

behavior ratings and a correlation of conceptual association ratings may share the same conceptual schema. In fact, multisource feedback ratings may be relatively accurate in reflecting the raters' perception of the ratee's performance, and the normative structure raters may reflect a unique and different conceptual schema regarding the pairing of the competencies.

The third research question, "What are the relative magnitudes of the differences between the correlation matrices from different sources?" seeks to elucidate the absolute magnitude of the differences between the correlation matrices, both among the source types and between the normative and source type groups, by computing the Euclidean distance between each pair of correlation matrices. The results indicated that the absolute magnitude of the differences in correlations among the competency dimensions and source types used in this study is quite small, and the small absolute magnitude of differences in correlations among the competency dimensions and source types is consistent with previous multisource feedback assessment research which indicated that source differences typify multisource feedback data.

The sample of actual raters from EADS was among the first group of raters in the company using the newly introduced multisource feedback process for executive development. Therefore, those raters were inexperienced in the rating process, and that lack of familiarity with the process may have contributed to rating differences. Similarly, the raters did not receive rater training in frame of reference or behavioral observation to improve their accuracy (Woehr & Huffcutt, 1994) before they marked the feedback assessment form. Perhaps one explanation for source differences in the actual data is that raters may have spent more time and thought in deciding on a rating for aspects of

specific performance because they were using a “new” appraisal process. Had they been using an assessment process that they had previously used, they might have responded more indiscriminately. In the words of Aesop, “Familiarity breeds contempt.”

The absolute magnitude of the differences in correlations between the normative and source type inter-dimensional correlations is consistently large across all source types. As described earlier, previous research has found that actual data and the normative pattern may share the same conceptual schema. However, in this study the actual data and normative pattern differed. The sample of raters for the normative structure came from an industry background and experience similar to the EADS raters in the actual data. Further, for the past two years, the executives at Force Protection have used a new but different process from the EADS multisource instrument and multisource feedback assessment for performance evaluation purposes. Perhaps the Force Protection executives’ assessment experience and background better prepared them for rating a hypothetical questionnaire as compared to a rater who had not previously used a multisource feedback instrument. As a result, the pattern of their ratings reflected a unique conceptual schema related to performance dimensions. Previously discussed research into normative pattern structures used graduate research students who likely were less familiar with the processes they were being asked to rate.

In summary, one can conclude that the absolute magnitude of the differences in correlations among the competency dimensions and source types used in this study is quite small, and the absolute magnitude of the differences in correlations between the normative and source type inter-dimensional correlations is consistently large, indicating

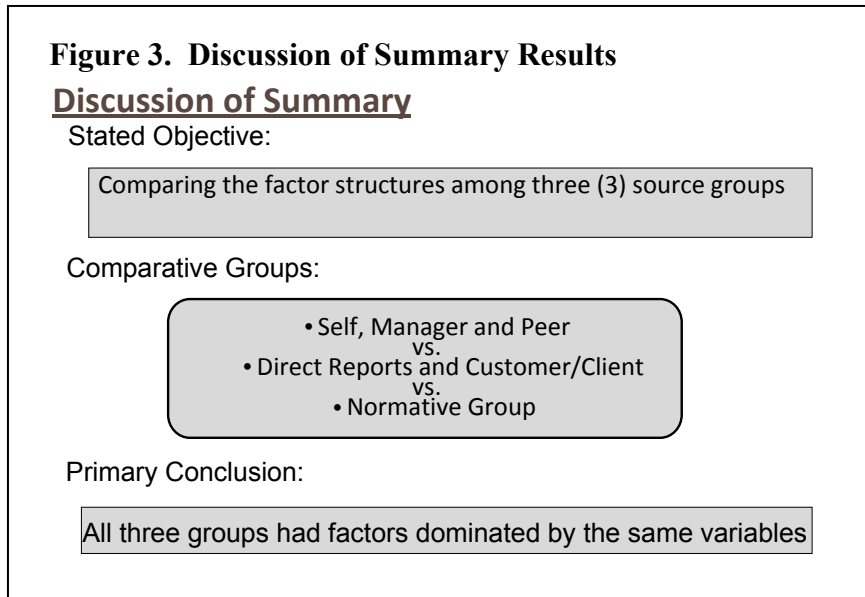
that the quantitative comparisons are in concurrence with the significant testing differences found in the results of the previous two research questions.

**Findings, Implications, and Limitations: Research Questions 4-6.** Question 4 asked, “How much difference do different implicit performance theories make in the weighted mean score that the ratees receive?” and examined the issue of whether differences between the raters’ policies of the different source types make a difference in the composite score across all the competency dimensions. Research hypotheses 16-25 explored the issue of whether differences between the rating policies of the different source types make a difference in the composite score across all the competency dimensions. The results of the correlational analyses were identical. The correlations between the composite scores (across the six competency dimensions) for each source type and the composite score using the regression weights of the other source types rounded to 1.0 in every case due to the near-equality of the means and standard deviations of the six competency dimensions within each source type. Thus, it can be concluded that there is insufficient evidence to support that differences between the rating policies of the various source types make a difference in the composite score across all the competency dimensions.

The tests of the differences in means were conducted using the paired t-test. The results indicated that there is sufficient evidence to support research hypotheses 26-33 and 35, that is, that there are significant differences in means. Only in the case of self weightings applied to customer/clients was there no significant difference in the means of the resulting scores.

Research question 4 was designed to build on Heslin's (2003) examination of implicit performance theories and the effect of prior judgment and implicit person theory on performance appraisals. His findings were consistent with those of Dweck (1999) and Aronson, Fried, and Good (2002): that induced incremental beliefs about implicit person theory can be sustained over a six-week period. Those findings provided Heslin reason to question the concept that an implicit person theory is a stable individual difference. Consistent with prior research, this research question was designed to review and interpret implicit theories of the five sources of the memory-based actual data across all the competency dimensions. Unfortunately, based on the results of this question, no conclusions can be drawn regarding how much difference the unique implicit performance theories make in the composite scores that the ratees received.

The fifth research question inquired as to the comparability between source types of the composition of the optimum number of rotated principal components computed for each source type's ratings of the six competency dimensions. The stated question was, "How do the structures of the matrices for various sources differ?" These comparisons consisted of pairing the rotated factors that were most similar between source types in terms of the competency dimensions that defined them and then computing the correlation between the rotated factor loadings for each pair of factors. Figure 3 condenses the summary results and the primary conclusion regarding the three comparative groups.



**Factor structures of self, manager, and peer ratings.** The factor structures of the self, manager, and peer ratings consisted of five factors that were identical across these three source types. The ratings of all three source types had factors dominated by the same variables. All three source types divided the Lead Courageously dimension between two factors dominated by the Drive Excellence and Innovation dimension and by the Communicate Honestly and Effectively dimension, respectively.

Wherry and Bartlett (1982) formulated more than 40 theorems regarding the relationship between the rater and the ratee. Theorem number 3 stated that rater accuracy will vary directly proportional to the relevancy of their prior contacts with the ratee. “Biases, by definition, grow from irrelevant contacts, while the true component of the rater's bias in either perception or recall can increase only as the true observations are more predominant than are irrelevant experiences” (Wherry, 1982, p. 532). This theorem

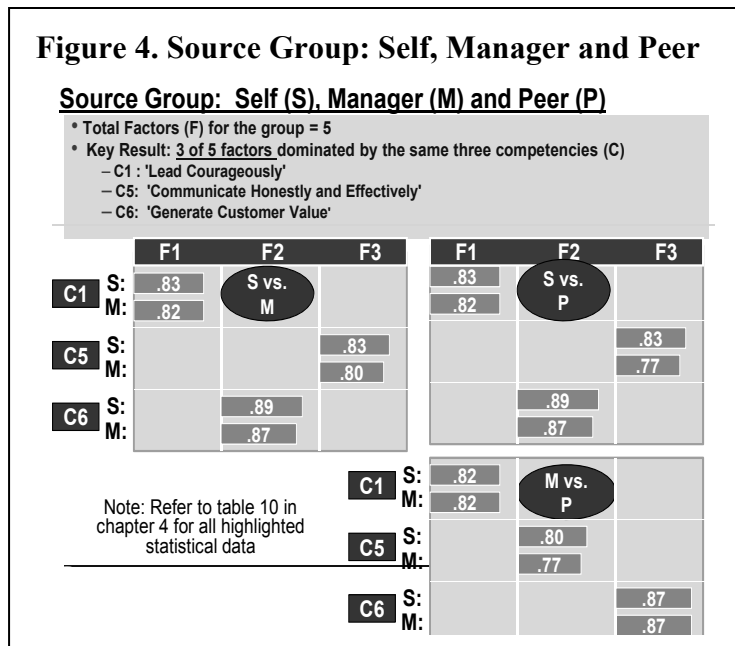
suggested that the closer a source type is to familiarity with the ratee's job, the more the source type is able to differentiate the competencies the job requires.

In the results of this research question, the grouping of ratee, peers, and manager ratings had five factors dominated by the same variables, and the five factors were identical across the three source types. The contacts between an executive ratee, his/her peers, and manager are frequent as are the activities they engage in together. For example, these executives attend the same strategy conferences and meetings and receive information to plan mutual objectives and goals that align with the corporation and with one another. Similarly, this grouping of executive, peers, and manager would work on projects and assignments together and attend social events related to those activities. The executive ratee, peers, and manager may experience similar information flow, receiving the same newspaper clippings regarding the company and industry, and the same emails from senior leadership containing company information. They may be part of the same network of close job associates and contacts. Wherry (1982) observed that those closer to the ratee on the job will more accurately rate the participant than those who are only casual acquaintances.

Three of the five factors were dominated by a single competency dimension, and the other two were dominated by two competency dimensions, one of which, in both cases, was Lead Courageously. Analyses involving the multisource feedback assessment of actual data source factors demonstrated that source factors for self, manager, and peer ratings had factors dominated by the same variables. All three source types divided the Lead Courageously dimension between two factors dominated by the

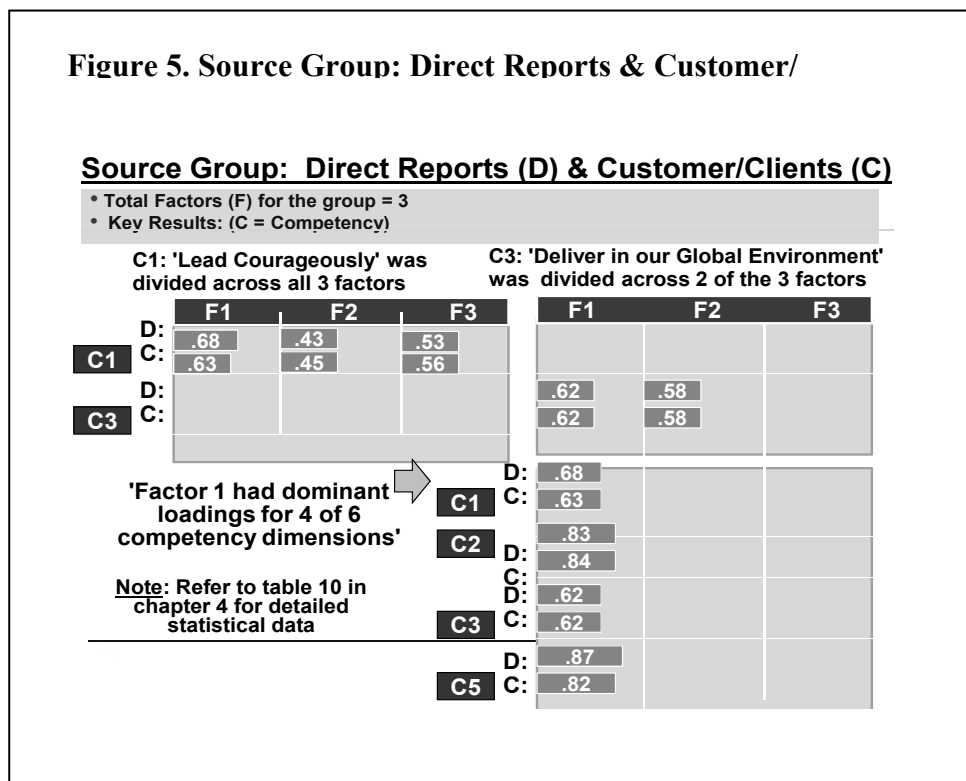
Drive Excellence and Innovation dimension and by the Communicate Honestly and Effectively dimension, respectively.

The results of this study indicate that the three source types divided the Lead Courageously dimension between two factors, dominated by the Drive Excellence and Innovation dimension and the Communicate Honestly and Effectively dimension. The linking of these competencies may support the notion of frequent interactions between the ratee, peers, and manager. The work of these positions tends to be more complex, and, as previously mentioned, the work relationships are interdependent. Consequently, the linking of a ratee, peers, and managers with the three competencies of Lead Courageously, Driving Excellence and Innovation, and Communicate Honestly and Effectively with these three source types may represent the “what” and “how” of organizational interaction that these executives share. Figure 4 illustrates that three of five factors were dominated by the same three factors.





**Factor structures of direct reports and customer/client ratings.** The factor structures of the direct reports and customer/clients consisted of only three factors, which were dominated by the same variables for both source types. The Lead Courageously dimension was divided across all three factors, and Deliver in our Global Environment was divided across two of the factors. One of the three factors contained the dominant loadings for four of the six competency dimensions. A second factor was dominated by three competency dimensions, two of which (Lead Courageously and Deliver in our Global Environment) were shared with the aforementioned factor. The third factor was dominated by two competency dimensions, one of which was Lead Courageously, and the other, Drive Excellence and Innovation. Figure 5 illustrates the findings for direct reports and customer/clients.



For direct reports and customer/clients, this converse of the aforementioned theorem by Wherry (1982) regarding rater proximity suggests that the further a source type is from direct familiarity with the ratee's job, the less the source type is able to differentiate the competencies the job requires. Although the contacts between an executive ratee, his/her direct reports, and customer/clients may also occur frequently, the relationship between ratee, direct reports, and customer/clients may have more physical and psychological distance than the relationship between ratee, manager, and peers. In a global company such as EADS, where the manager and subordinate, as well as the customer/client, are often in different locations, distance between raters and ratee may also preclude everyday interaction. The ratee, direct reports, and customer/clients would most likely not attend the same strategy conferences and meetings. Further, work on projects and assignments together and social interactions would tend to be less ad hoc, thus creating a psychological distance. Similarly, the information flow for direct reports and customer/clients might be received and/or filtered through others as opposed to receiving it "first hand" in meetings or emails.

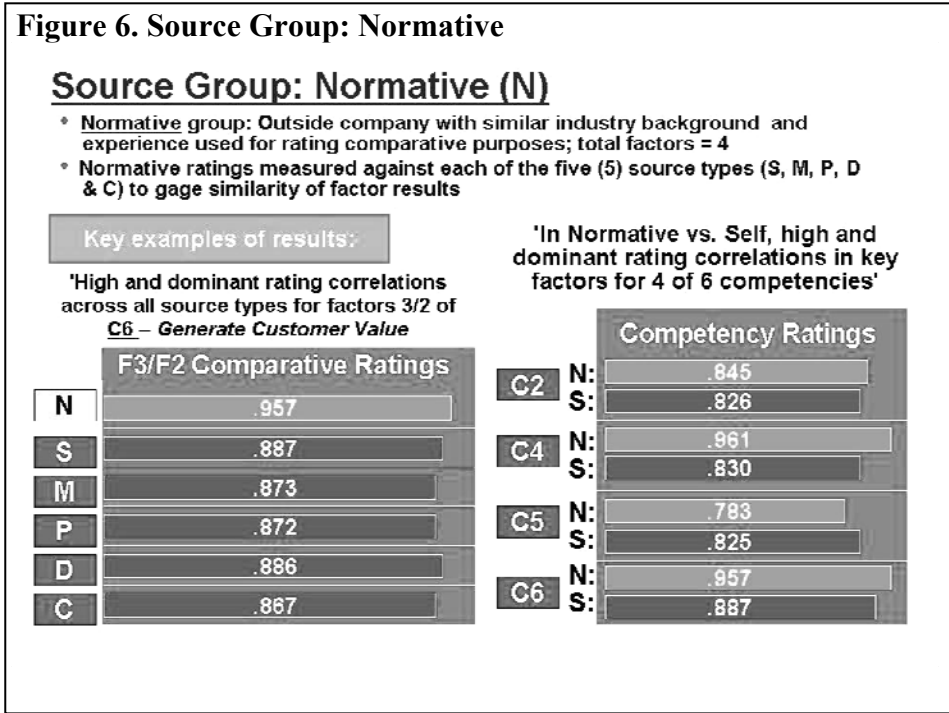
The results for research question 5 indicated that for direct reports and customer/clients, the Lead Courageously dimension and the Deliver in our Global Environment dimension were dominant competencies. The direct reports and customer/clients may evaluate the ratee on his/her ability to align goals and objectives to strategy and keep commitments (Lead Courageously). Also, in a global company such as EADS, working effectively across cultural and geographic boundaries (Deliver in our Global Environment) appears to be a competency that is essential to both direct reports and customer/clients when evaluating the ratee.

In summary, the results of research question 5 may indicate that the meaning of the source factor within multisource feedback assessments differs, depending on the source. The self, manager, and direct reports link the competencies of Lead Courageously, Drive Excellence and Innovation, and Communicate Honestly and Effectively, while direct reports and customer/clients link Lead Courageously with Deliver in our Global Environment.

The key purpose for the use of multisource feedback at EADS is leadership development. Thus, providing feedback to the ratee of how differently he/she is viewed by the manager, peers, direct reports, and customer/clients can help shape the subsequent coaching and developmental planning activities. The ratee may also compare his/her self-awareness with the different source views, and, with reflection, ultimately make needed behavioral changes.

**The factor structure of the normative group.** The sixth and final research question inquired as to the similarity of the rotated factor structure to the similarly rotated factor structure of the normative group. Ratings of each of the five source types are derived from actual data. Stated, the question is, “How similar is the rotated factor structure (maximum number of non-error principal components) to the similarly rotated factor structure of the normative group?” As with previous questions, these comparisons consisted of pairing up the rotated factors that were most similar between the normative factor solution and that of each source type in terms of the competency dimensions that defined them, and then computing the correlation between the rotated factor loadings for each pair of factors. The normative factor structure differed both from the five-factor structure of the self, manager, and peer raters and from

the three-factor structure of the direct report and customer/client raters. The normative factor structure consisted of four factors.



As illustrated in Figure 6, the normative group saw Deliver in our Global Environment as part of the same underlying construct as Develop Self and Others, whereas the self, manager, and peer raters saw it as representing a completely separate construct. The raters who completed the questionnaire which formulated the normative factor solution comprised the senior leadership team at Force Protection. They are part of the same industry as the leaders of EADS, the raters of the actual memory-based ratings. Consequently, the resulting factor structure of the normative group, though different from the five-factor structure of the self, manager, and peer raters and the three-factor structure of the direct reports and customer/client raters may represent a conceptual rating schema

from individuals who have similar backgrounds and experiences as the EADS raters. Similarly, both groups of raters are familiar with the multisource feedback assessments and the actual job duties of the ratees so that they could conceptualize the competencies being rated within an aerospace work environment. This parallel of background and experience may account for the fact that the only factor on which the normative group and the self, manager, and peer groups fully agreed as to its composition was the one defined by Generate Customer Value. Meeting the needs and expectations of customer/clients is a common shared value of executives whose strategy, goals, objectives, and rewards structure is determined by customer/client satisfaction with both their relationships and products.

The normative group saw Deliver in our Global Environment as part of the same underlying construct as Develop Self and Others, perhaps as a consequence of the type of work performed at Force Protection, Inc. As previously described, Force Protection, Inc. provides survivability solutions to support the armed forces of the United States and its allies (Force Protection Web site, 2011). Force Protection designs, manufactures, tests, delivers and supports its blast-and ballistic-protected products to increase the survivability of the users of the products. Its specialty vehicles are designed to protect their occupants from landmines, hostile fire, and improvised explosive devices. The company also has a vehicle support facility in Kuwait and provides military operations support in Afghanistan and Iraq. The leaders who completed the questionnaires that were used to form the normative structure have the responsibility to recruit, retain, develop, and care for direct reports, peers, managers, and themselves, often in a hostile environment. Therefore, the conceptual schema of these raters may link the Global

Environment of their own work milieu with the competency of Develop Self and Others in a different way than the EADS raters, who work globally but not solely in the “line of fire.”

Another aspect of Develop Self and Others is coaching for effective results. Giving Feedback and creating a climate for performance feedback is challenging in a global environment where geographical and psychological distance creates barriers that rarely exist when the leadership team co-exists in the same time zone and proximity.

In defining a common underlying construct, the normative group also saw Lead Courageously to be strongly linked with Communicate Honestly and Effectively. In contrast, the self, manager, and peer raters saw Lead Courageously to be almost evenly split between the Communicate factor and the Drive Excellence and Innovation factor. The competency Lead Courageously may ask for the ratee to be effective in clarifying roles and responsibilities for a diverse work environment. If so, then communicating this information clearly and allowing for dialogue may be particularly important in an environment marked by geographic distances. Conceptual schema governing the ratings by the normative sample may have included thinking about diverse backgrounds, workplace diversity, and ensuring organizational alignment of those different elements. If the aforesaid pattern of thinking existed, it might explain the linking of Lead Courageously and Communicate Honestly and Effectively in a common underlying construct. The ratings by the other two rating source type groups, direct reports and customer/clients, produced factors which, although highly similar between the two source type groups, departed quite substantially from those of the normative group.

The one additional factor that did not emerge in either the direct report or customer/client source type group was different for each group. The direct reports did not see the factor that the normative group defined in terms of Develop Self and Others and Deliver in our Global Environment dimensions. Instead, they included these dimensions in a highly general factor that also included Lead Courageously, Drive Excellence and Innovation, and Communicate Honestly and Effectively. This difference suggests that the normative group's implicit theories in linking the dimensions of Develop Self and Others and Deliver in our Global Environment may be due to the focus on and the constant pressure of the responsibility for keeping people out of "harm's way" in the war zones of Afghanistan and Iraq. Again, these results may indicate that direct reports, while "looking upwards" at their management, may link with the manager's effectiveness in aligning work with strategy, goals, and objectives. Management's ability to take care of the team is dependent on the specific manager's ability to clearly communicate the information "downward" across cultural and geographic barriers, and facilitate open dialogue when ambiguity arises.

The customer/client source group included the Develop Self and Others and Deliver in our Global Environment dimensions on the same factor as did the normative group, but also diverged from the normative group by including in the same factor the Lead Courageously and Communicate Honestly and Effectively dimensions. Again, the organizational level of customer/clients may perceive the ratee responsible for strategy alignment, meeting commitments to delivery of quality products and services, on time, within budget, and information flow; whereas, the normative group has conceptual

schema focused on products, services, and people because of their unusual responsibilities.

The customer/client source group did not perceive the factor that the normative group defined exclusively in terms of Lead Courageously and Communicate Honestly and Effectively. The normative group's visualization of pairing of the dimensions of Lead Courageously and Communicate Honestly and Effectively suggests that aligning the strategic vision and communicating that vision may be competencies that are valued for executive development at Force Protection, Inc.

Whereas the normative group defined one factor exclusively in terms of the Generate Customer Value dimension, the customer/client source type group saw this factor as a much more complex combination of dimensions that additionally included Deliver in our Global Environment dimension and lesser amounts of two other dimensions. Perhaps this difference reflects the customer/client base that each group serves. The normative group's customer/clients are military service personnel who deal with life and death situations daily. Therefore, executive attention may be put on serving the needs of that customer/client. The actual data group's customer/clients are varied with both commercial and military systems needs, and the perceptions of achieving customer/client satisfaction and understanding global trends may mirror that complexity.

Both direct reports and customer/clients defined the structure of the factor dominated by Drive Excellence and Innovation most similarly to the normative group, but the former two groups also perceived Lead Courageously as a noticeable influence in the underlying construct. This inclusion of Lead Courageously, influencing Drive Excellence and Innovation for the direct reports and customer/clients, suggests that these



raters may have had more seniority in their current positions, enabling them to conceptually embed innovation and excellence into their strategic imperatives. In contrast, the normative group is a newer executive team that may be focused on delivering innovation and excellence in their current product mix and has not yet had the time to link it conceptually to a longer range strategic plan. The research findings, which examine the actual data set that now provides historical rater patterns, and the normative structure pattern that now provides rating information, not only may guide rates, facilitators, and coaches toward developmental opportunities, but also may help ratees better understand where they fit into the rater pattern. They can then decide if a rating is “cultural” or personal.

In summary, results of this study seem to indicate that the organizational level of the rater for the actual data influences the grouping of the dimensions and provides distinct differences in how the ratee is perceived. Similarly, the normative group was able to differentiate the ratings into four factors with the structures of the factors differentiating from the actual data raters. Thus, they refuted the notion of a systematic distortion hypothesis which characterizes ratings as an illusion in which “propositions about language” are confused with “propositions about the world” (D’Andrade, 1965, p. 215), and similarity in meaning may be confused for the likelihood of co-occurrence.

As previously discussed, Woehr, Day, Arthur, and Bedeian (1996) formulated a hypothesis in which ratings are systematically distorted toward preexisting concepts. The authors’ research suggested that systematic distortion schemas are more likely to be used when raters lack knowledge about a particular job or when there is a delay between the observations of performance and the ratings (Woehr et al., 1996, p. 418). In the present

study, the two groups (normative and actual raters) seem to be able to differentiate dimensions within the factor structures, and each uses unique, not systemically distorted, conceptual schemas to formulate their ratings.

This study contributes to the previous research on multisource feedback by representing the differences in rater sources and the differences between the normative structure and the rater sources and then by interpreting the possible implications of the differences.

### **Limitations**

Although this study adds to the body of literature regarding multisource feedback, it has limitations to consider. One such limitation is that once the questionnaire was distributed to Force Protection participants, the researcher could not clarify the items even if there were questions about them from respondents. Further, the EADS multisource feedback instrument is based on PROFILOR® but nonetheless is a customized instrument and, perhaps, unique. Though the questionnaire used in the present study resembles to a large extent the PROFILOR® multisource feedback instruments, future studies may consider repeating the study with other multisource instruments.

In this study, rating scales were used in the actual data to depict raters' perceptions regarding the ratee's performance on the six competencies described in the actual data. A possible limitation regarding the use of these rating scales is that raters might have been confused by the rating scale and, thus, answered the questionnaire inaccurately. To alleviate that possibility, a 6-point scale was utilized in the actual data multisource instrument. Typically, multisource feedback assessment instruments use a 6-

point scale that includes a choice of Don't Know (DK) or Not Applicable (NA) to eliminate the possibility of the rater's selection of the middle or average rating (Thiry, 2009).

Another limitation is the lack of rater training. Raters for the present study had no training in using the instrument except for online instructions. EADS is a global company and the decision to use online instructions instead of training may be cost-based. Similarly, the geography and language obstacles associated with rater training may have discouraged its inclusion in the process. Working with the external vendor, perhaps the sponsor of the process could incorporate some computer based training into the process to improve rater accuracy (Craig, Hayes, Preston, & Lebow, 2006). Additionally, translation of the EADS instrument from English to other languages may have complicated the ratings for those whose native language is not English.

In the actual data sample, participants picked their raters, giving rise to the potential of affecting rating accuracy because of an association between rater and ratee. Still another limitation is that the subject of performance appraisal or assessment could have raised concerns or suspicions about the instrument. Some responses may also have been based in emotion, thereby creating a potential threat to validity.

Additionally, the questionnaire distributed to Force Protection executives was limited in its ability to inquire as to how the ratings were judged or what influenced the rating decisions (Gall et al., 2003). Another consideration is that the sample used for the normative data comprised executives from Force Protection, a company from the same industrial base as EADS. This similarity may have affected the comparison between the normative structure and the actual data and, thus, replication of this study with other

normative structure samples may be clearly warranted. Regardless of these limitations, the research results will be useful to diverse organizations that use or will use multisource feedback.

### **Implications and Recommendations for Further Research**

Waldman, Atwater, and Antonioni (1998) noted that there is a lack of knowledge “on how or even whether 360 feedback really works” (p. 89). The present study is significant to the field of research regarding multisource feedback because it compiled source-specific feedback and demonstrated that when the results were displayed by source, the meanings of the competency ratings by source can be interpreted. Hoffman’s findings in 2006 prompted him to encourage future researchers to further examine the causes of multisource feedback source effects so that the continued use of multisource feedback processes would be valuable to participants and to organizations.

It would be feasible to adapt the multisource feedback report given to ratees so that the participant, facilitator, and coach would have source specific information to reflect upon and subsequently to use for developmental planning. Future research might examine an evaluation process that compares the source specific feedback information from the ratee, facilitator, and coach to feedback without source specific information. This type of evaluation might also assess whether or not the source specific information is useful for reflection, coaching, and planning. As previously discussed, the results of this study diverge from previous studies in other disciplines that utilized a normative structure to compare to actual data. A future study might replicate prior research that compared a normative structure with actual data by utilizing graduate research students as the sample with actual multisource data. Future studies might consider replicating this

study using another sample from the same industry for the normative structure to see if the results are similar or diverse. Future researchers might also choose to consider using a sample for the normative structure from a different industry from whence came the actual data. Finally, future researchers might wish to administer the conceptual questionnaire to the raters who generated the actual data sample and compare the resulting normative structure data to the actual source data and report on findings.

### **Summary**

Drawing on prior research of multisource feedback assessment and related fields, the present study endeavored to explore whether multisource feedback results were an accurate reflection of how raters perceive the ratee's performance and how their scores matched with the scores derived from a normative pattern structure. Previous research has encouraged continued research about source differences (Borman, 1997; Farr, 2006; Hoffman, 2006; Thiry, 2009; Waldschmidt, 2006). Consequently, this study examined the nature and interpretation of the meaning of source differences. First, this study sought to clarify whether there are differences between various sources. The results revealed significant differences between every source pair of sources and the sources and the normative structure. Additionally, the amount of difference was measured using Euclidian distance, and the results indicated that the absolute magnitude of the differences in correlations among the competency dimensions used in this study were small.

Next, the study sought to determine whether, for each source type, there were differences in the correlations between, and the means of, weighted means and the predicted scores based on the other source types' regression equations. The analyses

concluded that there was insufficient evidence to attribute the differences to the effect of different rating source policies. However, the test of the differences in means was conducted using the paired t-test and the analysis indicated that the differences, though small, were statistically significant, signifying that different implicit performance theories might make a difference in the composite scores that ratees receive.

Finally, the study sought to determine how the structures of the matrices for various sources differ and whether or not the combined data for all the rating source types more closely correspond to the factor structure of the normative group than any source type's structure. The results present evidence that the structure of the various sources differ and that there are specific differences between the normative factor structure and the source type structures for self, peers, and manager and for the grouping of direct reports and customer/clients. The present study supported the premise that raters may have unique cognitive framework that allows for rating differences to be demonstrated and then interpreted.

Thus, based on these results, researchers and organizational leaders have the potential to positively influence the interpretation of rater feedback from multisource performance feedback processes, resulting in feedback that is more impactful and accurate for ratees, facilitators, and coaches. This feedback may then be relied upon for development planning and implementation. The results that compared archival data with normative pattern may stimulate thought in leaders and encourage them to be cognizant of and attuned to the design, interpretation, and implementation processes for multisource feedback within their organizations.

## References

- Alimo-Metcalfe, B. (1998). Professional forum: 360 degree feedback and leadership development. *International Journal of Selection and Assessment*, 6(1), 35-44.
- Anderson, B. (1966). *The psychology experiment: An introduction to the scientific method*. Belmont, CA: Wadsworth.
- Antonioni, D., & Park, H. (2001). The effects of personality similarity on peer ratings of contextual work behaviors. *Journal of Applied Psychology*, 54, 331-360.
- Aronson, J., Fried, C. B. and Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, 38, 113-125.
- Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology*, 77, 836-874.
- Bailey, C., & Austin, M. (2006). 360 degree feedback and developmental outcomes: The role of feedback characteristics, self-efficacy and importance of feedback dimensions to focal managers' current role. *International Journal of Selection and Assessment*, 14(1), 51-66.
- Bales, R. F. (1970). *Personality and interpersonal behavior*. New York: Holt, Rinehart and Winston.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Bartram, D. (2004). Assessment in organizations. *Applied Psychology: An International Review*, 53(2), 237-259.

- Beauvois, J. L., & Dubois, N. (2000). Affordances in social judgment: Experimental proof of why it is a mistake to ignore how others behave towards a target and look solely at how the target behaves. *Swiss Journal of Psychology*, 59 (1), 16-33.
- Benedict, M. E., & Levine, E. L. (1988). Delay and distortion: Tacit influences on performance appraisal effectiveness. *Journal of Applied Psychology*, 73, 507-514.
- Bernardin H.J., and Beatty, R.W. (1984). *Performance appraisal: Assessing human behavior at work..* Boston, MA: Kent.
- Bernardin, H. J., Dahmus, S. A., & Redmon, G. (1993). Attitudes of first-line supervisors toward subordinate appraisals. *Human Resource Management*, 32, 315-324.
- Bogdan, R. C., & Biklen, S. K. (2003). Ethical issues in qualitative and quantitative research. *Qualitative Research for Education: An Introduction to Theories and Methods* ( 4<sup>th</sup> ed.). Boston: Allyn and Bacon.
- Boote, D. N., & Beile, P. (2005 Aug/Sept). Scholars before researchers: On the centrality of the dissertation literature. *Educational Researcher*, 34(6), 3-15; ProQuest Psychology Journals.
- Bordens, K. S., & Abbott, B. B. (2008). *Research design and methods: A process approach* (8<sup>th</sup> ed.), Boston: McGraw-Hill.
- Borman, W. C. (1974). The rating of individuals in organizations: An alternative approach. *Organizational Behavior and Human Performance*, 12, 105-124.
- Borman, W. C. (1983). Implications of personality theory and research for the rating of work performance in organizations. In F. J. Landy, S. Zedeck, & J. C. Cleveland



(Eds.), *Performance measurement and theory* (pp. 127-172). Hillsdale, NJ: Erlbaum.

- Borman, W. C. (1987). Personal constructs, performance schemata, and “folk theories” of subordinate effectiveness: Explorations in an army officer sample. *Organizational Behavior and Human Decision Processes*, 40, 307-322.
- Borman, W. C. (1997). 360 degree ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review*, 7(3), 299-315.
- Bozeman, Dennis P. (1997). Interrater agreement in multisource performance appraisal: A commentary. *Journal of Organizational Behavior*, 18(4), 313-316. doi: 10.1002/(SICI)1099-1379(199707)18:4<313::AID-JOB844>3.0.CO;2-O  
Retrieved from <http://www.jstor.org/stable/3100178>
- Bracken, D. (1994, September). Straight talk about multirater feedback, *Training and Development*, 44-51.
- Bradley, T. P. (2004, May). Analysis of leadership perceptions using multirater feedback (Doctoral dissertation, University of North Texas). *ProQuest Information and Learning Company*, 1-130. (UMI No. 3126557)
- Breslin, J. W. (1991). Breaking away from subtle biases in *Negotiation Theory and Practice*, eds. J. William Breslin and Jeffery Z. Rubin, (Cambridge: The Program on Negotiation at Harvard Law School, 1991), 247-250. Retrieved from <http://www.colorado.edu/conflict/peace/example/bres75>
- Bretz, Jr., R. D., Milkovich, G. T., & Read, W. (1992). *The current state of performance appraisal research and practice: Concerns, directions, and implications*. Center

for Advanced Human Resource Studies, Working Paper Series, Cornell University ILR School.

- Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54*, 106-148.
- Brutus, S.; Fleenor, J. W.; & London, M. (1997). Elements of effective 360-degree feedback. In Tornow, W., & London, M. (Eds), *Maximizing the Value of 360-degree Feedback: A Process for Successful Individual and Organizational Development* (pp. 11-27). San Francisco: Jossey-Bass.
- Camerer, C. (1988). Illusory correlations in perceptions and predictions of organizational traits. *Journal of Behavioral Decision Making, 1*, 77-94.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.
- Campbell, D. J., & Lee, C. (1988). Self-appraisal in performance evaluation: Development versus evaluation. *Academy of Management Review, 13*, 302-314.
- Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills: Sage.
- Center for Creative Leadership Web Site (2011). Retrieved from <http://www.ccl.org>
- Chapman, L. J. (1967, February). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior, 6*(1), 151-155.
- Chapman, L. J., & Chapman, J. P. (1967, June). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72*(3), 193-204.
- Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*(3), 271-280.

- Cheung, G. W. (1999). Multifaceted conceptions of self-other disagreement. *Personnel Psychology, 52*, 1-36.
- Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1-27.
- Chory, R. M., & Westerman, C. Y. (2009). Feedback and fairness: The relationship between negative performance feedback and organizational justice. *Western Journal of Communication, 73*(2), 157-181.
- Conway, J. M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management, 22*, 139–162.
- Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self ratings. *Human Performance, 10*, 331–360.
- Cooper, W. H. (1981, September). Halo affect and rating/rater differences. *Psychological Bulletin, 90*(2), 218-244.
- Cooper, W. H. (1983). Internal homogeneity, descriptiveness and halo: Resurrecting some answers and questions about the structure of job performance rating categories. *Personnel Psychology, 36*, 489-502.
- Coovert, M. D., Craiger, J. P., & Teachout, M. S. (1997). Effectiveness of the direct product versus confirmatory factor model for reflecting the structure of multimethod-multirater job performance data. *Journal of Applied Psychology, 82*, 271–280.

- Cote, J. A., & Buckley, M. R. (1987, August). Estimating trait, method, and error variance: generalizing across seventy construct validation studies. *Journal of Marketing Research*, 26, 315-319.
- Craig, S.B.; Hayes, B.C.; Preston, L.A.; & Lebow, R. (2006, April). Rater training for 360 assessment: Creating beta and gamma change. In B.C. Hayes (Chair), *The Four "Rs" of 360 Feedback: Determinants of its Effectiveness*. Symposium presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.
- Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64(3), 391-418.
- Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Crotty, M. J. (1998). *The foundation of social research: Meaning and perspective in the research process*. Thousand Oaks, CA: Sage Publications, Inc.
- D'Andrade, R. G. (1965). Trait psychology and componential analysis. *American Anthropologist*, 67(5), 215-228.
- Dahlstrom, W. G., & Welsh, G. S. (1960). *Conceptual similarity of an MMPI handbook: A guide to use in clinical practice and research*. Minneapolis: University of Minnesota Press.
- Dalessio, A. T. (1998). Using multisource feedback for employee development and personnel decisions. In J. W. Smither (Ed.), *Performance appraisal: State of the art in practice* (pp. 278-327). San Francisco: Jossey-Bass.

- Dalton, M. (1996). Multi-rater feedback and conditions for change. *Consulting Psychology Journal: Practice and Research*, 48(1), 12-16.
- DeVellis, R. F. (1991). *Scale development: theory and applications*. Newbury Park, CA: Sage.
- Diefendorff, J., Silverman, S., & Greguras, G. (2005, April). Measurement equivalence and multisource ratings for non-managerial positions: Recommendations for research and practice. *Journal of Business and Psychology*, 19(3), 399-425.
- Dugan, K. W. (1989). Ability and effort attributions: Do they affect how managers communicate performance feedback information? *Academy of Management Journal*, 32, 87-114.
- Dweck, C. S. (1999). *Self-Theories: Their role in motivation, personality, and development*. Philadelphia: The Psychology Press.
- Dweck, C. S., Chiu, C., & Hong, Y. (1995). Implicit theories and their role in judgments and reactions: A world from two perspectives. *Psychological Inquiry*, 6.
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256-273.
- EADS web site (2011). Retrieved from <http://www.eads.com>
- Facteau, C. L., Facteau, J. D., Schoel, L. C., Russell, J. E., & Poteet, M. L. (1991). Reactions of leaders to 360-degree feedback from subordinates and peers. *Leadership Quarterly*, 9, 427-448.
- Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology*, 86, 215-227.

- Farh, J. L., & Werbel, J. D. (1986). Effects of purpose of the appraisal and expectation of validation on self-appraisal leniency. *Journal of Applied Psychology, 71*, 527-529.
- Farh, J. L., Werbel, J. D., & Bedeian, A. G. (1988). An empirical investigation of self appraisal-based performance evaluation. *Personnel Psychology, 41*, 141-156.
- Farr, J. L. (2006). In D. W. Bracken & C. W. Timmreck (co-chairs). *The Evolution of 360 Feedback: What Hath We Wrought?* Symposium at the 21st annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.
- Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*, 127-148.
- Ferris, G. R., Yates, V. L., Gilmore, D. C., & Rowland, K. M. (1985). The influence of subordinate age on performance ratings and causal attributions. *Personnel Psychology, 38*, 545-557.
- Fiske, D. W. (1982). Convergent-discriminant validation in measurements and research strategies. In D. W. Fiske, D. Brinberg, & L. H. Kidder. *New directions for methodology of social and behavioral science, 12. Forms of Validity in Research*, (pp. 77-92), San Francisco: Josey-Bass.
- Fleenor, J. W., & Prince, J. M. (1997). *Using 360 degree feedback in organizations: An annotated bibliography*. Greensboro, NC: Center for Creative Leadership.
- Fletcher, C., Baldry, C., & Cunningham-Snell, N. (1998). The psychometric properties of 360 degree feedback: an empirical study and a cautionary tale. *International Journal of Selection and Assessment, 6*, 19-25.

- Force Protection Web Site (2011). Retrieved from <http://www.forceprotection.net>
- Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th ed.). New York: McGraw-Hill.
- Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*(1), 75-90.
- Funder, D. C., Parke, R. D., Tomlinson-Keasey, C., & Widaman, K. F. (Eds.) (1993), *Studying lives through time: Approaches to personality and development*. Washington, DC: American Psychological Association.
- Gall, M. D., Gall, J. P., & Borg, W. R. (2003). *Educational research: An introduction* (7<sup>th</sup> ed.). Boston: Pearson.
- Geddes, D. (1993). Examining the dimensionality of performance feedback messages: Source and recipient perceptions of influence attempts. *Communication Studies*, *44*, 200- 215.
- Gliner, J. A., Morgan, G. A., & Leech, N. L. (2009). *Research methods in applied settings: an integrated approach to design and analysis*. New York and London: Routledge, Taylor & Francis Group.
- Goffin, R. D., & Jackson, D. N. (1992). Analysis of multitrait multirater performance appraisal data: Composite direct product method versus confirmatory factor analysis. *Multivariate Behavioral Research*, *27*, 363-385.
- Greenberg, J. (1986a). The distributive justice of organizational performance appraisals, In H.W. Bierhoff, R.L. Cohen, & J. Greenberg (Eds.), *Justice in social relations*, (pp. 337-352). NY: Plenum Press.

- Greenberg, J. (1986b). Determinants of perceived fairness of performance evaluations, *Journal of Applied Psychology*, 71, 340-342.
- Greguras, G. J. (2005, Spring). Managerial experience and the measurement equivalence of performance ratings. *Journal of Business and Psychology*, 19(3). doi: 10.1007/s10869-004-2234-y
- Gruenfeld, L. W., & Weissenberg, P. (1966). Supervisory characteristics and attitudes toward performance appraisals. *Personnel Psychology*, 19, 143-151. doi: 10.1111/j.1744-6570.1966.tb02024.x
- Guion, R. M. (1965). *Personnel testing*. New York, NY: McGraw-Hill.
- Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, 11, 385-398.
- Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41, 41-62.
- Hauenstein, N. M. A., & Alexander, R. A. (1991). Rating ability in performance judgments: The joint influence of implicit theories and intelligence. *Organizational Behavior and Human Decision Processes*, 50, 300-323.
- 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.
- Heisler, W. J. (1996). 360-degree feedback: An integrated perspective. *Career Development International*, 1(3), 20-23.



- Hellqvist, N. (2008). *Global integration of performance management within MNC's. A theoretical framework*. Unpublished manuscript, University of Vaasa, Finland.
- Heslin, P. A. (2003). *The effect of prior judgements and implicit person theory on performance appraisals*. (Doctoral dissertation). Joseph L. Rotman School of Management, University of Toronto. ISBN: 0-612-84768-3.
- Heslin, P. A., Latham, G. P., & VandeWalle, D. (2005). The effect of implicit person theory on performance appraisals. *Journal of Applied Psychology, 90*(5). doi: 10.1037/0021-9010.90.5.842
- Heslin, P. A., & VandeWalle, D. (2008). Managers' Implicit Assumptions about Personnel. *Association for Psychological Science, 17*(3), 219.
- Hoffman, B. J. (2006, August). *Disentangling the meaning of multisource feedback: An examination of the nomological network surrounding source and dimension factors*. (Doctoral dissertation, The University of Tennessee). UMI Number: 3235481
- Hogan, E. A. (1987). Effects of prior expectations on performance ratings: A longitudinal study. *Academy of Management Journal, 30*, 354-368.
- Hooper, M. A. (2004). Using the Baldrige criteria to strengthen educational leadership systems: the development and validation of a 360-degree feedback instrument. *Dissertations Abstracts International*. (UMI No. 3138750)
- Huber, V. L., Neale, M. A., & Northcraft, G. B. (1987). Judgment by heuristics: Effects of ratee and rater characteristics and performance standards on performance-related judgments. *Organizational Behavior and Human Decision Processes, 40*, 149-169.

- Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (pp. 5, 141-197), Greenwich, CT: JAI Press.
- Jennrich, R. J. (1970, June). An asymptotic chi<sup>2</sup> test for the equality of two correlation matrices. *Journal of the American Statistical Association*, 65(330), 904-912.
- Johnson, R. B., & Christensen, L. B. (2000). *Educational research: Quantitative and qualitative approaches*. Boston: Allyn and Bacon.
- Jones, J., & Bearley, W. (1996). 360 degree feedback: strategies, tactics and techniques for developing leaders. Amherst, MA: HRD Press.
- Kane, J. S. (2004). *Performance distribution assessment: A new method for measuring and managing performance*. Paper published on PDA Web Site. Retrieved from <http://www.assess.co.nz/pages/PDAPaper.doc>
- Kenney, R. A., Schwartz-Kenney, B. M., & Blascovich, J. (1996): Implicit leadership theories: Defining leaders described as worthy of influence. *Personality & Social Psychology Bulletin*, 22 (11), 1128-1143.
- Kerlinger, F. N. (1986). *Foundations of behavioral research* (3<sup>rd</sup> ed.). New York: Holt, Rinehart & Winston.
- Kozlowski, S. W., & Kirsch, M. (1987). The systematic distortion hypothesis, halo, and accuracy: An individual-level analysis. *Journal of Applied Psychology*, 72, 252-261.
- Kruglanski, A. W., & Ajzen, I. (1983), Bias and error in human judgment. *European Journal of Social Psychology*, 13, 1-44

- Lance, C. E., & Bennett, W. (1997). *Rater source differences in cognitive representations of performance information*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, St. Louis, MO.
- Lance, C. E., Hoffman, B. J., Baranik, L. E., & Gentry, W. A. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review, 18*, 223-232.
- Lance, C. E.; Teachout, M. S.; & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology, 77*, 437-452.
- Lance, C. E., & Woehr, D. J. (1989). The validity of performance judgments: Normative accuracy model versus ecological perspectives. In D. F. Ray (Ed.), *Southern Management Association Proceedings* (pp. 115-117). Oxford, MS: Southern Management Association.
- Lance, C. E.; Woehr, D. J.; & Fiscaro, S. A. (1991, January). Cognitive categorization processes in performance evaluation: Confirmatory tests of two models. *Journal of Organizational Behavior, 12*(1), 1-20.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72-107.
- 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. (1999). A Research Agenda for the Situational Interview. In G. R. Ferris & R.W. Eder (Eds.), *The employment interview: Theory, research and practice*, (pp. 169-182). Newbury Park, CA: Sage.
- Latham, G. P., & Wexley, K. N. (1994). *Increasing productivity through performance appraisal*. Reading, MA: Addison-Wesley.
- Lawler, E. E. (1967). The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology*, 51, 369-381.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563-575.
- Leadership Practices Inventory Web site (2011). Retrieved from [http://www.lpionline.com/lpi\\_360.html](http://www.lpionline.com/lpi_360.html)
- Lepsinger, R., & Lucia, A. (1997). *The art and science of 360 feedback*. San Francisco: Jossey-Bass Pfeiffer.
- Leslie, J. B., & Van Velsor, E. (1995). *A look at derailment today: North America and Europe*. Greensboro, NC: Center for Creative Leadership.
- Liff, J. (2010). *Expanding the social-cognitive framework: understanding the role of implicit person theories in a complex task feedback environment*. (Doctoral dissertation, Colorado State University).
- Lin, Y. (2007). Cultural implications of self-other agreement in multisource feedback: comparing samples from U.S., China, and globally dispersed teams. *Dissertations Abstracts International*, 1-XXX (UMI No. 3138750)
- Lingle, J. H., & Ostrom, T. M. (1979). Retrieval selectivity in memory based impression judgments. *Journal of Personality and Social Psychology*, 37, 180-194.

- Lingle, J. H., Geva, N., Ostrom, T. M., Leippe, M. R., & Baumgardner, M. H. (1979). Thematic effects of personal judgments on impression formation. *Journal of Personality and Social Psychology*, 37, 674-687.
- Lodzinski, A. C. (1991). *The systematic distortion hypothesis and diagnostically meaningful student ratings of university instructor teaching effectiveness* (Doctoral dissertation, University of Windsor, Ontario, Canada). ISSN 0-315-72792-6.
- Lomakina, N. V. (2008). *Predicted adjustment and its relationship to 360-degree behavioral ratings and personality measures in leaders considered for expatriate assignments* (Doctoral dissertation, Walden University College of Social and Behavioral Sciences).
- 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 Resources Management*, 32(2,3), 353-372.
- London, M., & Smither, J. W. (1995). Can multisource feedback change perceptions of goal accomplishment, self-evaluation, and performance-related outcomes? Theory based applications and directions for research. *Personnel Psychology*, 48, 803-839.
- Longnecker, C. O., & Goff, S. J. (1990). Why performance appraisals still fail. *Journal of Compensation and Benefits*, 6(3), 36-41.
- Lord, R. G., Foti, R. J., & Phillips, J. S. (1982). A theory of leadership categorization. In J. G. Hunt, U. Sekaran, & C. Schriesheim (Eds.), *Leadership: Beyond*

*establishment views*, (pp. 104-121). Carbondale, IL: Southern Illinois University Press.

Lord, R., & Maher, K. (1993). *Leadership and information processing: Linking perceptions and performance*. Boston, MA: Unwin Hyman.

Matens, J. (1999). Conducting effective performance reviews. *Modern Casting*, 89(6), 54-46. Retrieved from <http://www.thefreelibrary.com/Conducting effective performance reviews.-a055112607>

Maurer, T. J., Barbeite, F. G., & Mitchell, D. R. (2002, March). Predictors of attitudes toward a 360-degree feedback system and involvement in post-feedback management development activity. *Journal of Occupational & Organizational Psychology*, 75(1), 87-107. Retrieved from [http://web.ebscohost.com/ehost/delivery? vid=18&sid=6c57b97d-3b94-4011-bb49 1-25](http://web.ebscohost.com/ehost/delivery?vid=18&sid=6c57b97d-3b94-4011-bb49-1-25)

Maurer, T. J.; Raju, N. S.; & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology*, 83, 693-702.

McCauley, C. D., & Moxley, Jr., R. S. (1996). Developmental 360: How feedback can make managers more effective. *Career Development International*, 1(3), 15-19.

McCauley, C. D., & Van Velsor, E. (2004). *The center for creative leadership handbook of leadership development*. San Francisco: John Wiley & Sons Jossey-Bass.

McDaniel, C., & Gates, R. (1991) *Contemporary marketing research*. St. Paul, MN: West.

McGarvey, R., & Smith, S. (1993). When workers rate the boss. *Training*, 30(3), 31-34.

- McGarvey, R., & Smith, S. (1997). When workers rate the boss. In Fleenor, J. W. & Prince, J. M. (Eds.), *Using 360 degree feedback in organizations: An annotated bibliography*. Greensboro, NC: Center for Creative Leadership.
- Mertens, D. M. (2005). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods* (2<sup>nd</sup> ed.). Thousand Oaks, CA: Sage Publications.
- Mohrman, Jr., A. M., Resnick-West, S. M., & Lawler III, E. E. (1990). In J. W. Fleenor & J. M. Prince (Eds.), *Using 360 degree feedback in organizations: An annotated bibliography*. Greensboro, NC: Center for Creative Leadership.
- Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater and level effects in 360 degree performance ratings. *Personnel Psychology*, *51*(3), 557-576.
- 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* (pp. 130–148). San Francisco: Jossey-Bass.
- Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873–900.
- Murphy, K. R.; Gannett, B. A.; Herr, B. M.; & Chen, J. A. (1986). Effects of subsequent performance on evaluations of previous performance. *Journal of Applied Psychology*, *71*(3), 427-431.

- Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review* 23(2).
- Newcomb, T. (1931). An experiment designed to test the validity of a rating technique *Journal of Education* 22, 279-289.
- Nowack, Kenneth M. (1993) 360-degree feedback: The whole story. In Fleenor, J. W. & J. M. Prince, J. M. (Eds.), *Using 360 degree feedback in organizations: an annotated bibliography*. Greensboro, NC: Center for Creative Leadership.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3<sup>rd</sup> ed.) New York: McGraw-Hill.
- Personnel Decisions International (2000). Unpublished research on PROFILOR® factor structure. Minneapolis, MN: Personnel Decisions International.
- Personnel Decisions International Web Site (2011). Retrieved from <http://www.personneldecisions.com>
- Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of job performance ratings: An examination of leader race, leader gender, and rater level effects. *Human Performance*, 9(2), 103-119.
- Raju, N. S.; Van der Linden, W.; & Fler, P. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement*, 19, 353-368.
- Reise, S. P.; Widaman, K.F.; & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114, 552-566.



- Rosti, Jr., R. T., & Shipper, F. (1998). A study of the impact of training in a management development program based on 360 feedback. *Salisbury Journal of Managerial Psychology, 13*(1/2), 77-89.
- Rowson, A. (1998). Using 360-Degree Feedback Instruments up, down, and around the world: Implications for global implementation and use of multirater feedback. *International Journal of Selection and Assessment, 6*(1), 45-48.
- Rungtusanatham, M. (1998, July). Let's not overlook content validity. *Decision Line, 10-13*.
- Sala, F. (2001). *It's lonely at the top: Executives' emotional intelligence self (mis)perceptions*. Consortium for Research on Emotional Intelligence in Organizations. Hay McBer. Retrieved from [http://www.eiconsortium.org/pdf/executive\\_emotional\\_intelligence\\_360](http://www.eiconsortium.org/pdf/executive_emotional_intelligence_360).
- Salam, S., Cox, J. F., & Sims, H. P. (1997). In the eye of the beholder: How leadership relates to 360-degree performance ratings. *Group & Organization Management, 22*, 185-209.
- Scullen, S. E.; Mount, M. K.; & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956-970.
- Scullen, S. E.; Mount, M. K.; & Judge, T. A. (2003). Evidence of the construct validity of developmental ratings of managerial performance, *Journal of Applied Psychology, 88*(1), 50-66. doi: 10.1037/0021-9010.88.1.50
- Sears, R. R.; Macoby, E. E.; & Levin, H. (1957). *Pattern of child rearing*, New York: Harper & Row.

- Shadish, W. R.; Cook, T. D.; & Campbell, D. T. (2002). *Experimental and quasi experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Shapira, Z., & Zevulun, E. (1989). Performance rating in organizations: a facet analysis interpretation of the multitrait-multirater approach. *Multivariate Behavioral Research*, 24, 209-232. Reprinted in *Megamot Journal of Behavioral Science*, 1991, 33, 477-491.
- Shaver, W. (1995, reprinted 2000). How to build and use a 360 degree feedback system. *Info-line*. The American Society for Training and Development (ASTD). Retrieved from <http://www.astd.org>
- Shepard, L.A. (1993). Evaluating Test Validity. In L. Darling-Hammon (Ed.), *Review of Research in Education*, 19. Washington, DC: AERA.
- Shweder, R. A. (1975). How relevant is an individual difference theory of personality? *Journal of Personality*, 43, 455-484.
- Shweder, R. A. (1977, December). Likeness and likelihood in everyday thought: Magical thinking in judgments about personality. *Current Anthropology*, 18(4), 637-657.
- Shweder, R. A., & D'Andrade, R. G. (1980). The systematic distortion hypothesis. In R. Shweder (Ed.), *New directions for methodology of social and behavior science*, 4, 37-58. San Francisco: Jossey-Bass.
- Singleton, R.A.; Straits, B.C.; & Straits, M.M. (1993). *Approaches to Social Research*. New York: Oxford University Press.

- Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: the role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, 52, 689-699.
- Smither, J. W., & Reilly, R. R. (1987). True intercorrelation among job components, time delay in rating, and rater intelligence as determinants of accuracy in performance ratings. *Organizational Behavior and Human Decision Processes*, 40, 369-391.
- Strull, T. K., & Wyer, Jr. R. S. (1989, January). Person memory and judgment, *Psychological Review*, 96, 58-83.
- Stajkovic, A., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin*, 124, 240-261.
- Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *Journal of Applied Psychology*, 74, 136-142.
- Sternberg, R. L.; Conway, B. E.; Ketron, J. L.; & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology*, 41, 37-55.
- Successful Manager's Leadership Program Web site (2011). Retrieved from [http://www.personneldecisions.com/Product\\_Service\\_Detail.aspx?Sub=Developing&ID=178](http://www.personneldecisions.com/Product_Service_Detail.aspx?Sub=Developing&ID=178)
- Tatsuoka, M., & Silver, P. (1988). Quantitative research methods in educational administration. In N. J. Boyar (Ed.), *Handbook of research on educational administration* (pp. 677-701). New York: Longman.

- Thiry, K. J. (2009). *Factors that affect peer rater accuracy in multirater feedback systems*. (Doctoral dissertation, Capella University). ProQuest LLC. (UMI 3366289)
- Tjosvold, D., & Halco, J. A. (1992). Performance appraisal of managers: Goal interdependence, ratings, and outcomes. *The Journal of Social Psychology, 132*, 629-639.
- Tornow, W. W. (1993). Perceptions or reality: Is multi perspective measurement a means or an end? *Human Resource Management, 32*,(2&3), 221-229.
- Triola, M. F. (2007) *Elementary statistics*, (10th ed.). MyStatLab Series. Boston: Addison Wesley.
- Tsui, A. S., & Ohlott, P. (1988). Multiple assessment of managerial effectiveness: Interrater agreement and consensus in effectiveness models. *Personnel Psychology, 41*, 779-803.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.
- Van Velsor, E., & Leslie, J. B. (1991). *Feedback to managers, volume II: A review and comparison of sixteen multi-rater feedback instruments*. Greensboro, NC: Center for Creative Leadership.
- Van Velsor, E., Leslie, J. B., & Fleenor, J. W. (1997). *Choosing 360: A guide to evaluating multirater feedback instruments for management development*. Greensboro, NC: Center for Creative Leadership Press.

- Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574.
- Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108–131.
- Waldman, D. A., Atwater, L. E., & Antonioni, D. (1998). Has 360 degree feedback gone amok? *Academy of Management Executive, 12*, 86-94.
- Waldschmidt, D. M. (2006). *Representing source perceptions in multisource feedback using structural equation modeling*. (Doctoral dissertation, Illinois Institute of Technology) (ISBN:9780542736414)
- Weatherly, L. A. (2004, May). Management by objectives. Society for Human Resource Management. Retrieved from [www.SHRM.org](http://www.SHRM.org)
- 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, Sr., R. J., & Bartlett, C. J. (1982, September). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35* (2), 521-551. First published online, 2006, December. doi: 10.1111/j.1744-6570.1982.tb02208.x
- Wilson, C. L. (2003). *How and why effective managers balance their skills*. Columbia, MD: Rockatech Multimedia Publishing.
- Wimer, S., & Nowack, K. M. (1998, May). 13 common mistakes using 360-degree feedback. *Training & Development, May*, 69-77.

- Woehr, D. J.; Day, D. V.; Arthur, W.; & Bedeian, A. G. (1996). The systematic distortion hypothesis: A confirmatory test of the implicit covariance and general impression models, *Basic and Applied Social Psychology*, 16(4), 417-434.
- Woehr, D. J., & Huffcutt A. I., (1994), Rater training for performance appraisal: A quantitative review, *Journal of Occupational & Organizational Psychology* 67, 3, pp. 189-205.
- Woehr, D. J.; Sheehan, M. K.; & Bennett, W. (2005). Assessing measurement equivalence across ratings sources: A multitrait-multirater approach. *Journal of Applied Psychology*, 90, 592–600.
- Wyer, R. S., & Srull, T. K. (1989). *Memory and cognition in its social context*. Hillsdale, NJ: Erlbaum.
- Yammarino, F. J., & Atwater, L. E. (1993, Summer-Autumn). Understanding self-perception accuracy: Implications for human resource management. *Human Resource Management*, 32(2-3), 231-247.
- Yammarino, F. J., & Atwater, L. E. (1997). Do managers see themselves as others see them? Implications of self-other rating agreement for human resources management. *Organizational Dynamics*, 25(4), 35-45.
- Yukl, G. (2006). *Leadership in organizations* (6th ed.). New Jersey: Prentice Hall.
- Zedeck, S., & Baker, H. T. (1972). Nursing performance as measured by behavioral expectation sales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance*, 7, 457–466.

Zentis, N. L. (2007). *The impact of 360 degree feedback on leadership development*  
(Doctoral dissertation, Capella University, 2007). ProQuest Information and  
Learning Company, 1-149 (UMI No. 3249898)

## Appendix A

### Performance Covariation Questionnaire

(Form A)

#### Directions

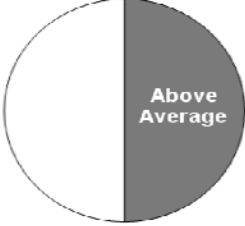
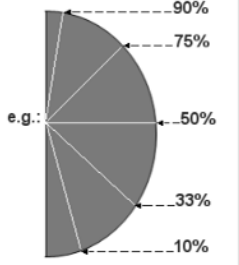
First, read the definitions of the 6 performance factors listed below until they are completely clear to you. Second, complete the ratings of the 15 pairs of performance factors that follow. In each case, narrow your focus to those managers who perform above average on the performance factor listed to left. Then decide what percentage of that subgroup, in your experience, also perform above average on the performance factor mentioned in the question to the right. Write in your percentage answer.

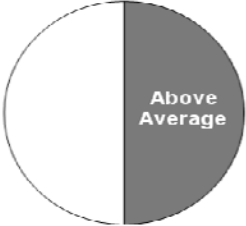
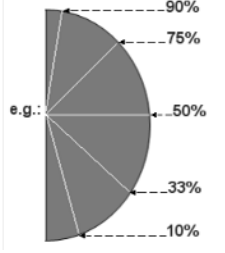
In making your ratings, please interpret “above average” to mean performing better than 50% or more of all managers in the comparison group.


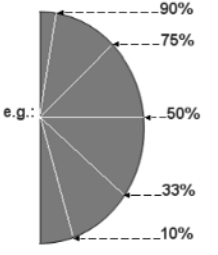
#### Performance Factor Definitions:

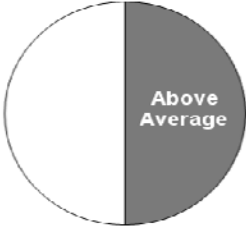
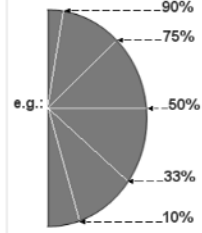
<b>Lead Courageously:</b> Build a vision and give clear strategic direction; make tough decisions and take responsibility for them; inspire and influence others; act with integrity	<b>Communicate Honestly &amp; Effectively:</b> Ensure direct and clear dialogue; proactively address conflicts and problems
<b>Drive Excellence &amp; Innovation:</b> Reliably deliver results by executing on time, cost, and quality; foster innovation, change, and continuous improvement	<b>Deliver in our Global Environment:</b> Demonstrate & apply international business knowledge; get things done through networks in & outside the company; be a team player across cultures and organizations
<b>Develop Yourself &amp; Others:</b> Challenge people to grow; actively develop people; learn continuously	<b>Generate Customer Value:</b> Focus the organization on customer needs and expectations; manage customer relationships


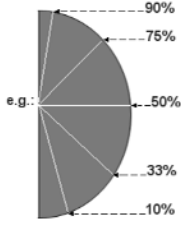


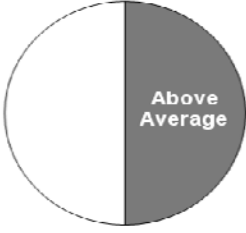
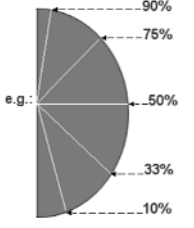
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b>Generate Customer Value</b> also perform above average on <b>Develop Yourself &amp; Others</b>?</p> 
<p><b>1. Generate Customer Value</b></p>	<p>_____ % (1) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b>Lead Courageously</b> also perform above average on <b>Drive Excellence &amp; Innovation</b>?</p>
<p><b>2. Lead Courageously</b></p>	<p>_____ % (2) (write in your answer)</p>


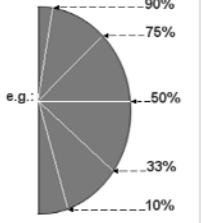
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <i>Develop Yourself &amp; Others</i> <u>also</u> perform above average on <i>Communicate Honestly &amp; Effectively</i>?</p> 
<p><b>3. Develop Yourself &amp; Others</b></p>	<p>_____ % (3) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <i>Deliver in our Global Environment</i> <u>also</u> perform above average on <i>Lead Courageously</i>?</p>
<p><b>4. Deliver in our Global Environment</b></p>	<p>_____ % (4) (write in your answer)</p>

<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b><i>Generate Customer Value</i></b> also perform above average on <b><i>Drive Excellence &amp; Innovation?</i></b></p> 
<p><b>5. Generate Customer Value</b></p>	<p>_____ % (5) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Lead Courageously</i></b> <u>also</u> perform above average on <b><i>Develop Yourself &amp; Others?</i></b></p>
<p><b>6. Lead Courageously</b></p>	<p>_____ % (6) (write in your answer)</p>

<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b><i>Drive Excellence &amp; Innovation</i></b> <u>also</u> perform above average on <b><i>Communicate Honestly &amp; Effectively?</i></b></p> 
<p><b>7. Drive Excellence &amp; Innovation</b></p>	<p>_____ % (7) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Develop Yourself &amp; Others</i></b> <u>also</u> perform above average on <b><i>Deliver in our Global Environment?</i></b></p>
<p><b>8. Develop Yourself &amp; Others</b></p>	<p>_____ % (8) (write in your answer)</p>

<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b><i>Communicate Honestly &amp; Effectively</i></b> also perform above average on <b><i>Lead Courageously</i></b>?</p> 
<p><b>9. Communicate Honestly &amp; Effectively</b></p>	<p>_____ % (9) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Deliver in our Global Environment</i></b> also perform above average on <b><i>Generate Customer Value</i></b>?</p>
<p><b>10. Deliver in our Global Environment</b></p>	<p>_____ % (10) (write in your answer)</p>

<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b><i>Drive Excellence &amp; Innovation</i></b> also perform above average on <b><i>Develop Yourself &amp; Others?</i></b></p> 
<p><b>11. Drive Excellence &amp; Innovation</b></p>	<p>_____ % (11) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Communicate Honestly &amp; Effectively</i></b> also perform above average on <b><i>Deliver in our Global Environment?</i></b></p>
<p><b>12. Communicate Honestly &amp; Effectively</b></p>	<p>_____ % (12) (write in your answer)</p>

<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b><i>Generate Customer Value</i></b> <u>also</u> perform above average on <b><i>Lead Courageously</i></b>?</p> 
<p><b>13. Generate Customer Value</b></p>	<p>_____ % (13) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Drive Excellence &amp; Innovation</i></b> <u>also</u> perform above average on <b><i>Deliver in our Global Environment?</i></b></p>
<p><b>14. Drive Excellence &amp; Innovation</b></p>	<p>_____ % (14) (write in your answer)</p>
<p><b>15. Communicate Honestly &amp; Effectively</b></p>	<p>_____ % (15) (write in your answer)</p>

## Appendix B

### Performance Covariation Questionnaire

#### (Form B)

#### Directions


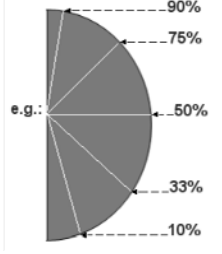
First, read the definitions of the 6 performance factors listed below until they are completely clear to you. Second, complete the ratings of the 15 pairs of performance factors that follow. In each case, narrow your focus to those managers who perform above average on the performance factor listed to left. Then decide what percentage of that subgroup, in your experience, also perform above average on the performance factor mentioned in the question to the right. Write in your percentage answer.

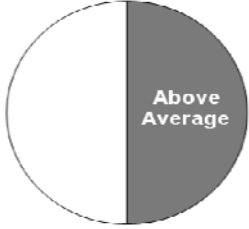
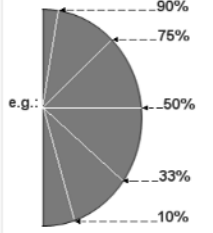
In making your ratings, please interpret “above average” to mean performing better than 50% or more of all managers in the comparison group.


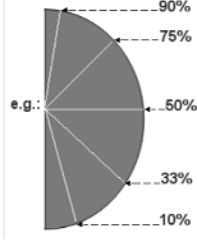
#### Performance Factor Definitions:


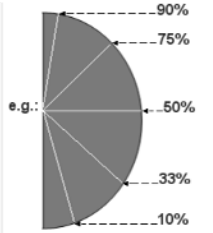
<b>Lead Courageously:</b> Build a vision and give clear strategic direction; make tough decisions and take responsibility for them; inspire and influence others; act with integrity	<b>Communicate Honestly &amp; Effectively:</b> Ensure direct and clear dialogue; proactively address conflicts and problems
<b>Drive Excellence &amp; Innovation:</b> Reliably deliver results by executing on time, cost, and quality; foster innovation, change, and continuous improvement	<b>Deliver in our Global Environment:</b> Demonstrate & apply international business knowledge; get things done through networks in & outside the company; be a team player across cultures and organizations
<b>Develop Yourself &amp; Others:</b> Challenge people to grow; actively develop people; learn continuously	<b>Generate Customer Value:</b> Focus the organization on customer needs and expectations; manage customer relationships

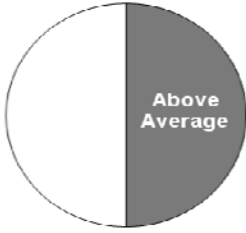
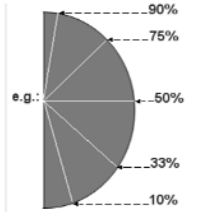


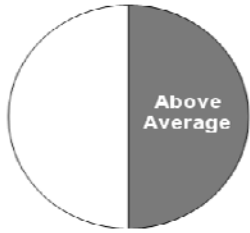
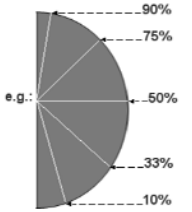
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b><i>Develop Yourself &amp; Others</i></b> also perform above average <b><i>Generate Customer Value</i></b> on?</p> 
<p><b>1. Develop Yourself &amp; Others</b></p>	<p>_____ % (1) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Drive Excellence &amp; Innovation</i></b> also perform above average on <b><i>Lead Courageously?</i></b></p>
<p><b>2. Drive Excellence &amp; Innovation</b></p>	<p>_____ % (2) (write in your answer)</p>


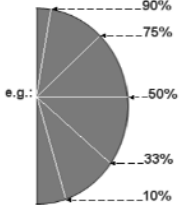
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b><i>Communicate Honestly &amp; Effectively</i></b> <u>also</u> perform above average on <b><i>Develop Yourself &amp; Others?</i></b></p> 
<p><b>3. Communicate Honestly &amp; Effectively</b></p>	<p>_____ % (3) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Lead Courageously</i></b> <u>also</u> perform above average on <b><i>Deliver in our Global Environment?</i></b></p>
<p><b>4. Lead Courageously</b></p>	<p>_____ % (4) (write in your answer)</p>

<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b><i>Drive Excellence &amp; Innovation</i></b> <u>also</u> perform above average on <b><i>Generate Customer Value</i></b>?</p> 
<p><b>5. Drive Excellence &amp; Innovation</b></p>	<p>_____ % (5) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Develop Yourself &amp; Others</i></b> <u>also</u> perform above average on <b><i>Lead Courageously</i></b>?</p>
<p><b>6. Develop Yourself &amp; Others</b></p>	<p>_____ % (6) (write in your answer)</p>

<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b><i>Communicate Honestly &amp; Effectively</i></b> also perform above average on <b><i>Drive Excellence &amp; Innovation?</i></b></p> 
<p><b>7. Communicate Honestly &amp; Effectively</b></p>	<p>_____ % (7) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Deliver in our Global Environment</i></b> also perform above average on <b><i>Develop Yourself &amp; Others?</i></b></p>
<p><b>8. Deliver in our Global Environment</b></p>	<p>_____ % (8) (write in your answer)</p>

<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b><i>Lead Courageously</i></b> <u>also</u> perform above average on <b><i>Communicate Honestly &amp; Effectively</i></b>?</p> 
<p><b>9. Lead Courageously</b></p>	<p>_____ % (9) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Generate Customer Value</i></b> <u>also</u> perform above average on <b><i>Deliver in our Global Environment</i></b>?</p>
<p><b>10. Generate Customer Value</b></p>	<p>_____ % (10) (write in your answer)</p>

<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <i>Develop Yourself &amp; Others</i> <u>also</u> perform above average on <i>Drive Excellence &amp; Innovation</i>?</p> 
<p><b>11. Develop Yourself &amp; Others</b></p>	<p>_____ % (11) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <i>Deliver in our Global Environment</i> <u>also</u> perform above average on <i>Communicate Honestly &amp; Effectively</i>?</p>
<p><b>12. Deliver in our Global Environment</b></p>	<p>_____ % (12) (write in your answer)</p>

<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p> 	<p>What percentage of these above-average performers on <b><i>Lead Courageously</i></b> <u>also</u> perform above average on <b><i>Generate Customer Value</i></b>?</p> 
<p><b>13. Lead Courageously</b></p>	<p>_____ % (13) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Deliver in our Global Environment</i></b> <u>also</u> perform above average on <b><i>Drive Excellence &amp; Innovation</i></b>?</p>
<p><b>14. Deliver in our Global Environment</b></p>	<p>_____ % (14) (write in your answer)</p>
<p>Consider the subgroup of managers who perform above average on the performance factor listed below.</p>	<p>What percentage of these above-average performers on <b><i>Generate Customer Value</i></b> <u>also</u> perform above average on <b><i>Communicate Honestly &amp; Effectively</i></b>?</p>
<p><b>15. Generate Customer Value</b></p>	<p>_____ % (15) (write in your answer)</p>

## Appendix C



Mr. Randy Hutcherson  
Chief Operating Officer  
Force Protection Industries, Inc.  
9801 Highway 78  
Ladson, SC 29456

Dr. Sue Kavli  
Dallas Baptist University  
Phone: (214) 333.5381  
Fax: (214) 333.6955  
E-mail: [suek@dbu.edu](mailto:suek@dbu.edu)

Please note that Margo Parker, Dallas Baptist University doctoral student, has permission of Force Protection to survey 32 senior leaders of Force Protection via written questionnaire. Ms. Parker will use these data as part of her research for her dissertation on multirater performance feedback at Dallas Baptist University and has permission to use the survey results now and in the future. Ms. Parker has agreed to provide a copy of the survey results to both Force Protection and Dallas Baptist University.


A handwritten signature in black ink, appearing to read "P. R. Hutcherson", is written over a horizontal line.

P. R. Hutcherson

November 10, 2010



## Appendix D

Dallas Baptist University		<b>Protection of Human Subjects Application</b>	
<b>Committee for the Protection of Human Subjects Action</b>			
Application is:		This Project's status is:	
<input checked="" type="checkbox"/> Approved		<input type="checkbox"/> Exempt	
<input type="checkbox"/> Approved with minor modifications (detailed below)		<input checked="" type="checkbox"/> Expedited	
<input type="checkbox"/> Denied		<input type="checkbox"/> Required Full Committee Review	
Official Date of Action:		4/25/11	
Comments:			
<b>Signature of Committee for the Protection of Human Subjects Chair</b>			
As Committee for Protection of Human Subjects Chair, I acknowledge that this research is in keeping with the standards set by the Dallas Baptist University and assure that the researcher has met all requirements for review and approval of this research.			
Signature of the Committee for Protection of Human Subjects Chairman:			
Signature: 		Date: 4/25/11	

## Appendix E

### Statistical Significance of Correlation Matrices

#### Question 1:

Are the differences between the sources' correlation matrices for the actual data statistically significant?

Compute:

Statistical test of differences

Between:

Corresponding correlations for @ pair of rating sources

Test by:

Computing differences matrix

Results (Differences)

Using:

Jennrich test

XXXX

Test Results (will show):

Leads to:

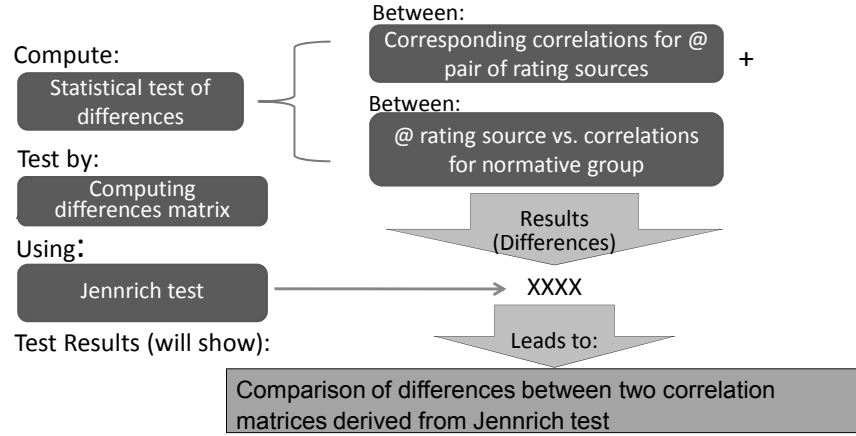
Comparison of differences between two correlation matrices derived from Jennrich test

## Appendix F

### Statistical Significance of Correlation Matrices

Question 2:

Are the differences between the sources' correlation matrices for the actual data source types vs. correlations for the normative group statistically significant?



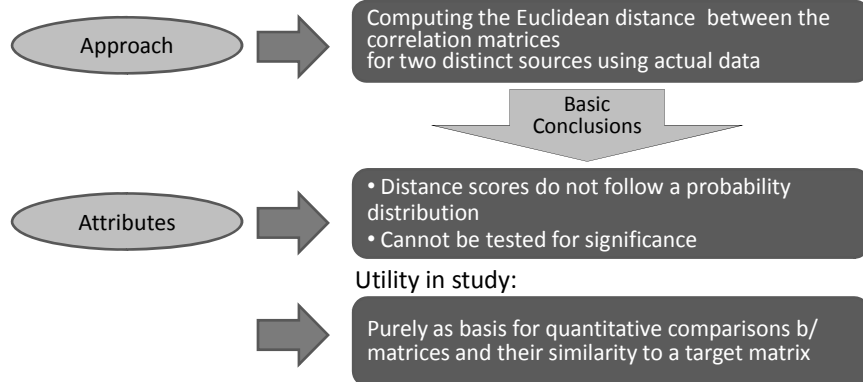
2

## Appendix G

### Correlation Matrices From Different Sources

#### Question 3:

What are the relative magnitudes of differences between the correlation matrices from different sources?



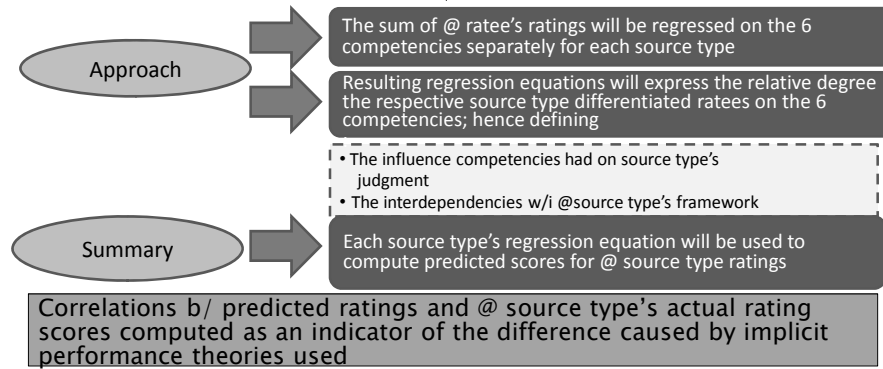
## Appendix H

### Impact of Implicit Performance Theories to Scoring

#### Primary Question 4:

How much difference do different implicit performance theories make in the total rating scores ratees receive?

- A. Use the differences in correlations b/ weighted mean scores for source types and predicted scores based on source types' regression equation?
- B. Use a T-test to achieve results for mean differences b/ @ source type's actual overall mean score and those w/ weights of @ other source type

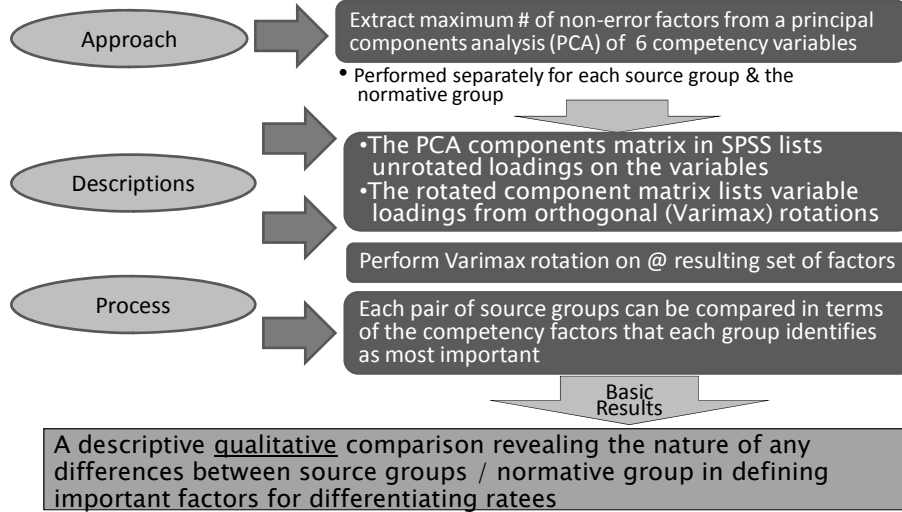


## Appendix I

### Differences in Source Structures

#### Question 5:

How do the structures of the matrices for various sources differ?



## Appendix J

### Rotated Factor Structure Similarities

#### Question 6:

How similar is the rotated factor structure (maximum # of non-error principal components) to the similarly rotated factor structure of the normative group?

