2012

MULTISOURCE FEEDBACK LEADERSHIP RATINGS: ANALYZING FOR MEASUREMENT INVARIANCE AND COMPARING RATER GROUP IMPLICIT LEADERSHIP THEORIES

Kim Gower
Virginia Commonwealth University

MULTISOURCE FEEDBACK LEADERSHIP RATINGS:

ANALYZING FOR MEASUREMENT INVARIANCE AND COMPARING RATER GROUP

IMPLICIT LEADERSHIP THEORIES

A dissertation proposal submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Business at Virginia Commonwealth University.

by

Kim Gower
B.S. Michigan Technological University, 1986
M.B.A. University of Michigan-Flint, 1995

Co-Director: Anson Seers
Professor, Department of Management
Virginia Commonwealth University

Co-Director: Margaret L. Williams
Interim Dean, School of Business Administration
Wayne State University

Virginia Commonwealth University
Richmond, Virginia
May, 2012

Dedication

To my students at:
Spring Arbor University
Virginia Commonwealth University
Virginia State University

Your passion and enthusiasm for learning made me pursue and persist.

To My Bonnie:

Your passion and enthusiasm for me kept me going all the way to the END!

Acknowledgements

It absolutely took a village, but here I am, baby!

To the VCU Management Department Administration and Faculty: thank you for giving freely of your time and talent along the way. Special thanks to Dr. Randy Barker, Dr. Glenn Gilbreath, Dr. Matt Rutherford, Dr. Jeffrey Krug, Dr. Randy Sleeth and Dr. Bob Andrews. Your personal and professional input made many days a little bit and even a lot bit easier.

To my academic friends who always said "You got this!" Thanks Dr. Barb Ritter, Dr. Pat Hedberg, Dr. Gayle Lawn-Day, Dr. Kiersten Maryott. To my Ph.D. peeps, ESPECIALLY to: Sheila, Tom, Ernest, Nicole, George, Bryan and Dave. To my friends and family who checked in regularly to ask if my "little story" (as my sister Jodi affectionately, or dismissively, referred to it) was finished: Nancy, Lise', Mom, Dad, Robin, Jodi, Craig, Spence, Chuck, Charlie and Leslie.

To my new home at Virginia State University: best assistant professor gig ever -The DEAN, Merri, Cheryl, Andrew, Mark, etc.

To my committee: Six years of thanks to Dr. Anson Seers, Dr. Mike McDaniel and Dr. Larry Williams – you have all gone above and beyond so many times to make me feel like one of the gang and to be my mentors and role models. Thanks especially, Dr. Seers, for always saying "you are making progress" even if I wasn't! To Dr. Jean Gasen and Dr. Mary Hermann– thank you for your positive and helpful input along the way. A special thanks to Dr. Bob Anderson, who is allowing me to use the data he collected from his successful leadership development company, The Leadership Circle (TLC), for my research.

Finally, to the two people with whom I would not be here, Dr. Jeff Pollack and Dr. Margaret Williams. JP, you are my constant, my kind shoulder and my listening ear, and when you say "How are you?" you really want to know. You are truly the best friend in the world and I am so lucky to call you mine. Peg, I told you many times I never would have made it without you; you thought I was being nice…I was not! Your door was always open, you gently (or not) prodded me when I needed it, and most of all you seemed to truly believe in me. My life has been made immeasurably better because you both took the time you did not have and the energy that you should not have had and gave them generously and graciously to me.

To all my village people…THANKS!

Table of Contents

List of Tables

Page

Abstract

MULTISOURCE FEEDBACK LEADERSHIP RATINGS: ANALYZING FOR MEASUREMENT INVARIANCE AND COMPARING RATER GROUP IMPLICIT THEORIES

By Kim Gower Ph.D.

A dissertation proposal submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Business at Virginia Commonwealth University.

Virginia Commonwealth University, 2012

Co-Director: Anson Seers
Professor, Department of Management
Virginia Commonwealth University

Co-Director: Margaret L. Williams
Interim Dean, School of Business Administration
Wayne State University

This research outlines a conceptual framework and data analysis process to examine multisource feedback (MSF) rater group differences from a leadership assessment survey, after testing the measures for equivalence. MSF gathers and compares ratings from supervisors, peer, followers and self and is the predominant leadership assessment tool in the United States. The results of MSF determine significant professional outcomes such as leadership development

opportunities, promotions and compensation. An underlying belief behind the extensive use of MSF is that each rater group has a different set of implicit leadership theories (ILTs) they use when assessing the leader, and therefore each group is able to contribute unique insight. If this is true, research findings would find rater group consistency in leadership assessment outcomes, but they do not. A review of group comparison research reveals that most empirical MSF studies fail to perform preliminary data exploration, employ consistent models or adequately test for measurement equivalence (ME); yet industry standards strongly suggest exploratory methods whenever data sets undergo changes, and misspecified models cause biased results. Finally, ME testing is critical to ascertain if rater groups have similar conceptualizations of the factors and items in an MSF survey. If conceptual ME is not established, substantive group comparisons cannot be made. This study draws on the extant MSF, ILT and ME literature and analyzes rater group data from a large, application-based MSF leadership database. After exploring the data and running the requisite MI tests, I found that the measures upheld measurement invariance and were suitable for group comparison. Additional MI tests for substantive hypotheses support found that significant mean differences did exist among certain rater groups and dimensions, but only direct report and peer groups were consistently significantly different in all four dimensions (analytical, interpersonal, courageous and leadership effectiveness). Additionally, the interpersonal dimension was the most highly correlated with leadership effectiveness in all five rater groups. The overall findings of this study address the importance of MSF data exploration, offer alternative explanations to the disparate leadership MSF research findings to date and question the application use of MSF tools in their current form.

Introduction

Multisource Feedback and Leadership Development

In the United States, Multisource Feedback (MSF) is the most widely used employee evaluation and development tool (Atwater, Waldman, Ostroff, Robie, & Johnson, 2005). The process of MSF - also known as multisource performance ratings (MSPRs) and 360 feedback - gathers and compares employee assessment ratings from multiple rater group observers including peers, subordinates, supervisors, and self (Viswesvaran, Schmidt, & Ones, 2002). The results are used to determine significant corporate investments in training and development programs, promotions and employee compensation (Scullen, Mount, & Judge, 2003).

MSF is also the predominant leadership development tool, and is based on the premise that rater groups differ in their perceptions of a leader's performance (Hoffman, Lance, Bynum, & Gentry, 2010). Each group's perceptual differences contribute unique information to the overall assessment of the leader, creating a more well-rounded analysis of the leader's strengths and opportunities for improvement. The rater group results are compared to the leader's own perception of performance to highlight areas where agreements and disparities exist between groups, and help with self-improvement to further develop her or his leadership skills and abilities (Hooijberg & Choi, 2000).

The scientific dimensions and theories underlying the usefulness of leadership MSF in applied settings is suggested to only be as strong as the rater groups' Implicit Leadership Theories (ILTs), or subjective evaluation of leadership performance (Scullen et al., 2003). From an early age, through experience and observation, humans cognitively categorize behaviors (Rosch, 1975) and these categorizations create abstract and subjective prototypes or "best

examples" of representatives of these categories (London & Smither, 1995; Offermann, Kennedy, & Wirtz,1994). Leadership prototypes form through exposure to different social and interpersonal events, and through observing those in leadership positions. Hence, we categorize certain behaviors and characteristics based on our personal perceived match between the person we are rating, and our pre-existing perceptions, or ILTs (Epitropaki & Martin, 2004).

Given that ILTs reflect the structure and content of an individual's cognitive categories and prototypes, they bring forth leadership prototypes with no conscious effort on the part of the rater (Lord, Foti, & DeVader, 1984). Therefore, ILTs are not representative of objective reality, but are simply perceptions or labels that individuals hold in their minds about what leadership "is" and who leaders "are", and then subsequently attribute to those they perceive as "leaders" (Epitropaki & Martin, 2004).

Multisource feedback ILT's, therefore, would be evoked based on each rater group's assumptions and expectations of the leader. That is, underlying a peer rating is what is common among peers, and the underlying construct among other groups is what is common among each of those groups (Viswesvaran et al., 2002). For example, peer prototypes are more heavily weighted towards the relationship factors of the leader, superior's prototypes are more interested in objective measures of performance factors and followers' prototypes are swayed by a variety of factors such as relational, managerial, and professional knowledge (Fox & Bizman, 1988).

These unique perceptions and contributions from each group are supposed to assist in establishing a more comprehensive view of the leader's strengths and opportunities for improvement. Previous studies conclude that feedback from subordinates improves leader performance (Atwater, Roush, Fischthal, 1995), and feedback from peers offers unique insight into the leader's teamwork and cooperation behavior dimensions (Scullen et al., 2003).

Additionally, low appraisal scores from one's boss makes managers more open to the feedback from other groups (Waldman & Atwater, 2001).

If ILTs play a significant role in how we categorize and rate leaders, and MSF ratings from different groups provide unique and meaningful information about the target based on group membership (LeBreton, Burgess, Kaiser, Atchley, James, 2003), there should be a consistent body of MSF leadership results, but there is not. Instead, results from MSF studies range from finding no differences in leadership perceptions and ratings between rater groups - supervisors, peers, and subordinates (Tsui & Ohlott, 1988) - to finding significant differences in ratings between these same groups (Gregarus & Robie, 1998).

A comprehensive review of the empirical and theoretical literature surrounding MSF reveals two consistent explanations for the differences. One is that even though there is general agreement that MSF is influenced by at least two overall factors, raters and dimensions, not all models in the literature appropriately represent these factors (e.g. Woehr, Sheehan, & Bennett, 2005; Hoffman et al., 2010). If ILT's are present in rater group perceptions, but the theoretical and perceptual dimensions of interest differ, then failure to model rater groups and properly identified dimensions in each MSF leadership study will result in biased end results due to mis-specified models (Hoffman, et al., 2010).

The second explanation is found when investigating the large amount of literature that explores the underlying methodological properties of group comparison ratings (Viswesvaran et al., 2002). Whenever studies employ cognitively based measures, such as MSF surveys, the appropriate methodological steps need to be taken to ensure that the respondents are measuring the same number and types of attributes (dimensions) across the source groups, and that the relation between specific items and a dimension are the same across rater groups (Woehr et al.,

2005). If these conditions are met, the measures in the survey instrument are conceptually equivalent and evoke the same conceptual frame of reference and magnitude between rater groups (Vandenberg, 2002).

These steps, known as tests for measurement equivalence/invariance (ME or MI), are necessary when using and comparing group measures because surveys requiring cognitive responses always carry a probability that rater groups have different perspectives. These differences might be apparent when interpreting the items in the measures, the measures themselves, or even the trait and behavior factors that evolve from the measures. Most group comparisons are done via analysis of variance (ANOVA) of the group means, or analyzing t-test statistics (t-test) which assesses if means between two groups are different. Neither of these tests the underlying assumptions of between group comparisons that each group equivalently assesses the latent variable, that each group equivalently associates the operationalizations and latent variables, and that there is equivalent assessment of the operationalizations to the same degree and other unique factors across groups (Vandenberg & Lance, 2000). An example might occur in the self-rater group, which is notorious for its inconsistencies (Conway, 1999). If self views strategic planning skills as paramount to be an effective leader, where a direct report views strategic planning skills as being secondary to good interpersonal skills, both rater groups see both factors as being important, but not to the same extent. Also, inside those factors, there could be items that self versus direct report do not view similarly. For instance, the item "I encourage others to lead" might be a behavior that is evident with self and peers, but not with direct reports. Now there could be a situation where self rates her or himself high and direct reports do not, but only because they do not recognize the behavior or do not interpret the item the same way as self. This item might also be a strong quality of self, but peer exhibits some bias in answering

due to the ramifications of the outcomes of MSF instruments. In the first scenario, self and direct report are not assigning the same weight to the factors. In the second situation, the item could be interpreted differently by self and direct report, and answered with bias by the peer.

In summary, with almost all published MSF results, mean differences and correlations between groups and factors are reported, all without the benefit of MI testing to see if the factors are being interpreted the same way and if items are being interpreted the same way, and/or if biases exist between rater groups. ME testing confirms that the measures being used are invariant, i.e. that the numeric values between groups are on the same measurement scale, and therefore the items on the measure and the constructs they measure are invariant across groups. This allows methodologically sound group comparisons of the MSF results to be made (e.g. Vandenberg, 2002; Vandenberg & Lance, 2000). If ME testing shows the measures are not invariant, then the groups had different conceptualizations of the measure and/or have perceived the items in the measure as relating to something different than what is being measured (Facteau & Craig, 2001). Given that MSF leadership assessments contain up to 130 items, up to 30 measures and up to 16 factors, the importance of testing the conceptual equivalency of the measures across groups is critical. A review of the existing MSF leadership studies show minimal or no attempts to test the data for MI. Failure to perform ME tests in group research renders any substantive between group comparisons inappropriate (Woehr et al., 2005).

This potential for "interpretational confounding" - differences in rater interpretation of a variable, such as performance, compared to the a priori assigned meaning of the variable by the researcher or differences in interpretations of variables between rater groups (Burt, 1976), clouds our ability to attribute meaning to structural relations among the constructs (Anderson & Gerbing, 1988). Unfortunately, most MSF studies do not test or satisfactorily test for

measurement equivalence among groups, and ME itself is misunderstood (Meade & Bauer, 2007) and inadequately applied (Vandenberg & Lance, 2000).

The role of ILT's in leadership MSF is based on perceptual differences among rater groups, too, exposing two areas for interpretational confounding – ILT's and MI. If the appropriate steps are not taken from a methodological standpoint to discern if the items and measures are conceptually invariant, then it is impossible to determine if group differences and similarities exist based on ILT's or measurement error. This further complicates the analysis of any results or conclusions with regards to dimensions or rater groups, and from a practical perspective should cause even more concern. Nearly all Fortune 500 companies use MSF to assess managerial performance (Cheung, 1999), yet the application of the results differ extensively. As example, in the leadership field the leadership "development" MSF surveys often result in an assessment of the person's leadership performance and/or effectiveness as well. In fact, over 90% of MSF results, regardless of what type of survey they are supposed to administering, are applied to significant human resource decisions (Greguras, Robie, Schleicher, & Goff, 2003). This makes it critical that a consistent body of research exists as to the adequacy of MSF leadership assessment tools, and that the application of the results of these tools are based on true group comparison results, not interpretational confounding.

There is also a sizeable investment in any applied MSF process, including purchasing the correct instrument and/or hiring the appropriate consultants, administering the survey to the rater groups, collecting and analyzing the data, and providing feedback. Moreover, the results of MSF lead to long term organization investments in coaching programs, pay raises and bonuses and executive personnel strategies. If the leadership assessment tool is not rigorously tested and validated then the investment in the system and the analysis of the group comparison

information for leader development and organization decision-making is subject to potentially grave and expensive errors.

The empirical discrepancies in MSF leadership field and the continuing arguments as to where and why these disagreements exist presents an excellent research opportunity, and there are three existing gaps this research aims to fill. First, although theoretically posited that MSF is valuable because different rating sources contribute unique perspectives to understanding performance (Borman, 1997), no empirical models exist in the literature using five levels of internal raters: self, subordinates, peers, supervisors, and supervisor's boss. The data for this study comes from a MSF leadership development database that has not been used in previous published research, and it includes five rater groups. Second, most existing empirical MSF leadership assessment studies have been conducted using the same database, with a very large percentage of the raters and leaders being white males (e.g.,Hoffman et al., 2010; Scullen et al., 2003). Research shows that an inherent bias in the leadership appraisal system, due to leadership prototypes invoking white male ILTs, has caused disparate advancement for subgroups (Roberson, Galvin, & Charles, 2007), that group similarity effects ratings (Rosette et al., 2008) and that individual differences, including gender and race, preclude raters from responding to cognitive measures in the same way (Vandenberg & Lance, 2000). As white males continue to decrease as a portion of the leadership and rater pool in a more heterogeneous workforce (Cascio & Aguinis, 2008), it is important to use and interpret group data from a rater sample that is more representative of today's workforce. This study includes more demographically representative data, with nearly half female respondents.

Finally, before making group comparison statements, the rater groups and dimensions must undergo MI testing. The existing MSF leadership studies, using the same database, have

not been validated outside of the database owner, and although the researchers often make changes in the item and dimension structures they do not use any data exploration processes, as recommended whenever changes are made (Field, 2011). There is also minimal MI testing present in the studies, and where it exists it uses old testing procedures that are now considered invalid, and/or uses sample sizes that are too small for appropriate MI testing. This study uses significant data exploration methods to review and validate the items and dimensions, and is the first study with a large enough sample in each of the five rater groups (minimum requirement of 600 respondents in each group [Vandenberg, 2002]) to conduct the appropriate ME tests.

This research proceeds as follows. First, I discuss the background and application of multi-source feedback as a significant part of the U.S. leadership assessment process. Next, I discuss implicit leadership theories and their proposed role in creating rater group differences, and summarize the existing MSF rater group research disparities. Then, I discuss measurement equivalence and its importance for group comparison research, review seminal MSF studies and propose my group comparison hypotheses. After describing the database and methods, I analyze my results and discuss how my findings make research and application contributions to the MSF leadership assessment field.

Multisource Feedback

Multisource Feedback ratings, or "360-degree" performance appraisal measures, compare ratings by multiple observers such as co-workers, subordinates, supervisors, and supervisor's bosses to self- ratings (Viswesvaran, et al., 2002). MSF began as a way to extend traditional performance appraisals (London & Smither, 1995) and now is an integral administrative decision-making and developmental tool in a majority of U.S. organizations (Roberson et al., 2007).

Administrative MSF ratings focus on assessing performance for making compensation, promotion and succession planning decisions (Borman, 1997). Although considered more lenient and less accurate than developmental ratings (Scullen et al., 2003), more than 90% of companies that use any type of MSF surveys apply the results administratively to a variety of critical human resource management areas. This makes MSF a significant part of the U.S. employee assessment process (Greguras, et al., 2003).

At the leadership level, MSF is the most popular development instrument in the US (e.g. Borman, 1997; Atwater et al., 2005). Feedback from subordinates improves leader performance (Atwater, et al., 1995), and improvements in leader performance impacts important employee outcomes such as improving subordinate job satisfaction and reducing subordinate intentions to leave (Atwater & Brett, 2006).

Unfortunately, at the MSF leadership level there is a muddled collection of results as to the differences between the rater groups, the correlations between the rater groups and even the dimensions of interest used in the surveys (Conway, 1999). In reviews of seminal work in the MSF leadership field there are models that exclude rater groups and results that feature large differences and no differences between rater groups and large and small correlations between

rater groups. In addition, there are a plethora of sub-dimensions, dimensions, and factors defining leadership skills and behaviors including interpersonal/human skills/human relations, administrative/technical skills/task-related, and leadership (management) competencies/effectiveness/development/performance (e.g. Conway, 1999; Hoffman et al., 2010; Mount et al., 1998; Scullen et al., 2003). This creates a murkiness surrounding management performance evaluations and leadership development assessments due to the lack of a clear nomological network and disagreement as to the purposes and applications of MSF leadership tools used in the field. Also, while there is a difference in MSF tools based on the purpose of the survey (administrative versus developmental) nearly all of the outcomes are used as performance evaluators (Gregarus et. al, 2003). In the leadership field itself, the most widely used set of instruments come from a self-described developmental database, but the outcomes are often projected as leadership performance results (Atwater, Ostroff, Yammarino, & Fleenor, 1998; Hoffman et al., 2010; Mount et al., 1993; Scullen et al., 2003), and most likely operationalized as such in the raters' minds.

The lack of clarity in defining a set of MSF leadership development factors, and a consistent set of items to include in those factors, makes for a wide range of reported outcomes in the field. The problem is compounded when mingling performance and developmental tools, creating an array of performance, competency, effectiveness and developmental items and dimensions, but trying to make them all "fit" into MSF leadership feedback research and appropriate application. If collected and evaluated correctly, MSF developmental leadership ratings are critically important for practitioners and researchers (Hooijberg & Choi, 2000). If raters from various organization levels do have different perspectives of the leader based on their group membership (Epitropaki & Martin, 2003), and these perspectives differ from the leaders'

own self-ratings (e.g., Scullen et al., 2003; Borman, 1997), then the time and capital investment spent finding, administering and using MSF leadership development tools is well worth the several billion dollar industry it has become.

A brief overview of the typical MSF rater groups and their proposed group ILT's, as well as a sample of some related research outcomes, is below.

Direct Reports/Followers/Subordinates

This group is posited to have the most complex relationship with leaders. Direct reports are concerned with fairness and leadership issues as well as their bosses' ability and willingness to aid in their own development (Scullen et al., 2003). Some authors suggest that only subordinates can truly assess leadership effectiveness because they are the only ones experiencing the full array of leadership characteristics such as managerial abilities, interpersonal relations, and professional knowledge (e.g., Beatty, 2005; Fox & Bizman, 1988; Kaiser, Hogan, & Craig, 2008). Other researchers conclude direct report ratings only correlate to the followers' self interest, such as leaders encouraging independent subordinate action receive higher subordinate ratings (Salam, Cox & Sims, 1997). Thus, some research finds that direct reports are an important source of MSF ratings, but other research finds that direct reports are the least reliable source (Barr & Raju, 2003).

Peers

Peer feedback is administratively valuable in making predictions of future performance and assisting in providing information for promotion decisions, salary allocation, and/or other administrative action because peers have better knowledge of the leader than other rater groups (Borman, 2005). Peers are more likely than other groups to depend on the interpersonal skills of the leader (Fox & Bizman, 1988), and peers offer a unique perspective based on teamwork and

cooperation that supervisors and followers might not experience (Scullen et al., 2003). In addition, peers stress compatibility (interpersonal relationships) with peers (Landy, Farr, Saal, & Freytag, 1976), and peer ratings are most reliable when assessing manager behaviors in common business situations (Greguras & Robie, 1998).

Other research, however, finds that peer ratings must be placed in context because they are susceptible to self-interest (Dierdorff & Surface, 2007), and peer feedback is not perceived by the leaders as useful (Facteau, Facteau, Schoel, Russell, & Poteet, 1998). In fact, leaders are least likely to recall feedback from peers (Smither, Brett, & Atwater, 2008).

Boss/Manager/Supervisor

The main advantage of collecting supervisor ratings (as compared to peer or subordinate ratings) is that supervisors typically have worked with many different subordinates and will make more "reasonably accurate" ratings of true performance based on past experience (Borman, 2005, p. 273). Bosses use more objective rationale than other groups, using quantitative measures such as budget and revenue (Scullen et al., 2003). Developmentally, leaders are more likely to recall feedback from supervisors than other raters (Smither, et al., 2008), and managers receiving lower appraisal scores from their bosses are more likely to perceive the upward feedback process to be useful (Waldman & Atwater, 2001).

When assessed in a practical performance setting, however, supervisors made the least accurate assessments (Gregarus & Robie, 1998), and although supervisors stress the importance of leader-subordinate interaction and initiative (Fox & Bizman, 1988), leaders encouraging independent subordinate action and challenging status quo receive lower ratings from their supervisors (Salam et al., 1997).

Supervisor's Boss

Interestingly, there appears to be no research regarding this group in the MSF literature although certain organizations, such as universities, rely heavily on this level of feedback for administrative and development purposes. This proposal hypothesizes that the supervisor's boss rater category will most closely model the supervisor rater category (closely enough to allow for aggregation of the two groups) and heavily emphasize quantitative qualities versus interpersonal qualities. This group should also offer the least complex rating for overall leadership effectiveness.

Self

Fundamentally, developmental MSF leadership assessment compares self-ratings to other ratings to evaluate the agreement level between self-perception and ratings from other groups (Atwater, et al., 1998). Although self is reported to "know the leader best" (Borman, 2005, p. 340), managers often lack insight into the impact of their behaviors (vanDierendonck, Haynes, Borrill, & Stride, 2007). Therefore, self-ratings provide a basis for the leaders to judge how they may have under- or overestimated their performance in each area (Borman, 2005). Much of the literature regarding self-evaluation agrees with Conway (1999) that self results are "puzzling".

It is apparent that different rater groups have different perspectives on leaders, and that existing research disagrees on rater group results in leadership MSF. The next section reviews ILTs and their proposed role in these MSF results

Implicit Theories and Leadership

Implicit theories evolved from the field of psychology and the theory of cognitive categorization. Beginning very early in the cognitive development process, through experience and observation, humans categorize behaviors. These categories are highly determined and include only items that are equivalent in the mind of the observer (Rosch, 1975). From these categories, we create abstract and subjective prototypes or "best examples" of representatives of these categories (e. g., London & Smither, 1995; Offermann et al., 1994). Therefore, when presented with a behavior that falls into one of our subjective categories we implicitly include the person demonstrating that behavior into our predetermined category.

Implicit theories of behavior categorization are triggered in two different ways, by cognitive factors and by performance cues (Bryman, 1987). Cognitive factor is based on object recognition (Rosch, 1975). As an example, if we perceive a similarity between the object and something we have observed before, we place that object in the same category. Performance cue is the cue validity of an entire category (Rosch, 1975). As an example, if we see a behavior or "performance" that fits one of our categories, we place the object in that existing category.

As humans, we use categorization for all cognitive processing, and in MSF research categorization plays a key role in how individuals appraise others' performance. An early summary of the MSF process defined appraisal as "the product of a set of social cognitive operations which includes acquisition of information through observation of performance, organization and storage of that information in memory, retrieval of information from memory, and its integration to form a judgment" (DeNisi, Cafferty, Meglino, 1984, p. 361). Even before that, researchers realized that although surveys of leadership were purported to be measuring

actual leadership behaviors, they were in fact measuring individuals' perceptions of leadership (Eden & Leviatan, 1976).

Implicit Leadership Theories, therefore, evolve not from the actual behaviors of the leader but from our exposure to different social and interpersonal events and observing those in leadership positions. From these we develop leadership prototypes (Epitropaki & Martin, 2004), which are individual cognitive categorization structures regarding the traits and behaviors that we expect from leaders based on our previous experiences (Offermann et al., 1990). So, when we are asked to rate a person in a leadership position, we select and organize information and categorize how the job should be performed (and by whom) based on our own leadership prototype (Lord et al., 1984). We use these theories/prototypes as a framework for evaluation so the ratings do not represent objective reality but rather subjective perceptions or labels that individuals place on leaders (Epitropaki & Martin, 2004).

A significant body of research posits that different employee rater groups hold different implicit leadership theories and perceptions of characteristics and performance because ILTs vary as a function of the context (group) within which an employee operates (Lord, Brown, Harvey, & Hall, 2001). It is also strongly suggested that different rater groups hold different self-interests and biases which might affect MSF leadership ratings (Tsui & Ohlott, 1984), further impacting and confounding rater group results. Current research indicates followers are interested in fairness and the leader's interest and ability to help them develop (Scullen et al., 2003), so this group emphasizes leadership abilities, interpersonal relationships and knowledge when evaluating their boss (Fox & Bizman, 1988). Peers show an interest in teamwork and cooperation and so they place the strongest emphasis on interpersonal relationships (Conway, 1999; Fox & Bizman, 1988), while supervisors are shown to be more interested in objective

measures of performance (Scullen et al., 2003). Finally, self-evaluations tend to be different than the other groups because individuals are not good at evaluating themselves (Harris & Schaubroeck, 1988).

In summary, ILTs are not representative of objective reality but rather perceptual abstracts or labels used to categorize individuals in leadership positions based on the rater's past experiences with other leaders, self-interests and individual needs (Epitropaki & Martin, 2004). In fact, past experiences with other leaders and even exposures to erroneous information about a leader effect raters' perceptions (Rush, Thomas, & Lord, 1977), and simply labeling somebody as a "leader" brings forth existing implicit leadership information to help the rater simplify and integrate the information they have about the person in question (Foti, Fraser, & Lord, 1982). Even raters sharing positive, self-perceived traits with a leader, such as both being "hard working", produces a more positive evaluation (Lord, Brown, & Freiberg, 1999).

Despite the premise that that MSF rater responses are influenced largely by the ILT's that raters "carry around" with them (Bryman, 1987, p. 130), and the large body of MSF research investigating differences in ILTs based on organization rater group membership (e.g. Tsui & Ohlott, 1988;) and individual idiosyncrasies (Mount, Judge, Scullen, Sytsma, & Hezlett, 1998), other MSF research disagrees. Support exists for ILT's being "stable trait-based stereotypes of leadership" across employee groups when it comes to perceptions of ideal leadership (Epitropaki & Martin, 2004, p. 308), and for a shared conceptualization of leadership behaviors based on group membership (Facteau & Craig, 2001). Yet another study suggests that ILT's are stable across organization members for task related behaviors, but not contextual behaviors (Johnson & Meade, 2010).In all, a spirited debated continues about the presence or absence of

ILT's in MSF surveys and, if present, at what rater level – individual or group, and what dimension – task-related, contextual and/or overall performance (effectiveness).

From an application standpoint, MSF systems are the primary tools for assessing leaders and continue to gain popularity. The fundamental theory behind MSF is that different rater sources contribute unique information to the process, producing a more well-rounded perspective of the leader's skills, behaviors and performance (LeBreton et al., 2003). It is not clear from the literature if this is true, and if it is, what causes the differences in ratings. This calls into question the fundamental underpinnings of MSF leadership assessment tools. The next section provides an overview of the areas of contention in the MSF literature, by raters and dimensions, and summarizes disparate results among studies.

Summary of MSF Rater Group and Dimension Disparities

The structure of performance ratings and their components continue receiving widespread empirical and theoretical attention due to the discrepancy in results (Hoffman et al., 2010). Rater groups (subordinates, peers, self, supervisor, and supervisor's boss) arguably have their own set of ILT's based on what they need and expect at their level, and therefore MSF should provide unique insights for leader development (e.g., Fox & Bizman, 1988; Johnson & Meade, 2010). Other research refutes the presence and/or impact of rater source group ILT's on MSF leadership ratings and finds that ILT's exist only at the individual rater level (Scullen et al., 2000). Similarly, studies of the measures typically included in MSF leadership development surveys – skill and task behaviors, interpersonal behaviors and general overall performance – offer a range of contrasting findings from the dimension level up to an overall performance global factor (e.g. Conway, 1999; Hoffman et al., 2010; Scullen et al., 2003; Viswesvaran et al., 2005).

The examples below highlight the contrasting findings at the rater source and dimension levels. A detailed outline of the seminal empirical studies in the MSF leadership assessment field is presented after the theoretical literature overview.

Rater Group

Direct Reports/Followers/Subordinates. One stream of literature states that subordinates are less experienced raters because leadership constructs develop and sharpen over time (Borman, 1987), and therefore direct reports are the least reliable sources of leadership ratings (Barr & Raju, 2003). Other research finds that subordinates provide significantly better feedback for leadership development than other groups, such as peers (Gregarus, et al., 2003), and that subordinates pay more attention to supervisor's behavior than peers (Maurer, Raju, & Collins, 1998).

Peers. Some research posits that peers, then subordinates and then supervisors are most reliable when rating practical situation performance by leaders (Greguras & Robie, 1998). Other research questions if peers feel they are in direct competition with the rating subject and therefore are more subjective when asked to provide feedback (Gregarus et al., 2003). Another study states that peers and supervisors have nearly complete between group agreement when assessing leaders (Viswesvaran et al., 2005), while another finds that peer and supervisor ratings are different (Pulakos et al., 1996).

Boss/Manager/Supervisor. Supervisors are sometimes reported to be the best source of ratings since they have more experience and are able to form a common rating perspective (Mount et al.,1998). Other research shows boss group ratings are significantly lower than other groups (Harris & Schaubroeck, 1988) and, when assessed in a practical performance setting, supervisors make the least accurate assessments (Gregarus & Robie, 1998).

Self. Self-ratings enjoy a spirited amount of research in many areas of management, and MSF is no different. Individuals are not good at evaluating themselves (Harris & Schaubroeck, 1998) and managers lack insight into their behavior impact (vanDierendonck et al., 2007), so it is theorized that MSF offers a way for leaders to gain a more objective view of their performance by comparing their results to their co-workers. On the other hand, some studies show that self-ratings are a reliable source (Barr & Raju, 2003) and that self-evaluation of ability may correspond closely to performance (Mabe & West, 1982).

To summarize, some MSF studies find that raters across all rater groups share a common conceptualization of managerial performance dimensions (e.g. Facteau & Craig, 2001; Mauer, et al., 1998; Scullen et al., 2003). Therefore, rater group source effects provide a small source of variance (Scullen et al., 2000; Woehr et al., 2005) and rater source effects should be

disconfirmed in MSF research (Viswesvaran et al., 2005). However, other studies find that rater groups hold different conceptualizations of performance (Lance & Bennett, 2000), and rater source effects provide a large source of variance (Hoffman, et al., 2010). The controversial findings in MSF research are also present when reviewing the performance dimension and factor results, too.

Dimensions and a General Performance Factor

There is no consistent framework for assessing leadership performance. In the current literature, anywhere from two to eight dimensions are tested, and often those are rolled into a single higher (general) performance factor for testing as well. The dimensions under review include task, technical and administrative skills, interpersonal, human skills and human relations behaviors, extra role behaviors and leadership skills (e.g. Hoffman et al., (2010); Mount et al., (1998); Scullen et al., (2003). Additionally, leadership MSF assessing military personnel include dimensions for following rules and military appearance (Woehr et al., 2005). Despite this range of dimensions, however, only two consistent factors emerge - task-oriented skills and interpersonal behaviors. As defined by Conway (1999) and refined by Scullen et al., (2003), task-oriented skills are related to the technical and administrative skills of the manager, like producing, planning and coordinating resources, and interpersonal behaviors are those that include leading and managing people in a variety of situations, as well as establishing positive working relationships. Here is where the MSF study results diverge again. For instance, one study finds rater groups share a common conceptualization of task-related and interpersonal skills and behaviors (Scullen et al., 2003) and another finds rater groups universally understand task behaviors but contextual behavior ratings are a product of individual rater implicit theories

(Johnson & Meade, 2010). Finally, a third finds rater groups differ in how they define skills and behaviors in both dimensions (Borman, 1997).

The higher order factor of overall leadership effectiveness/performance provides equally confounding findings. High consensus exists among raters when assessing a higher order dimension (e.g. Tsui & Ohlott, 1988; Viswesvaran, et al., 2005), and a low consensus exists among raters when assessing a higher order dimension (e.g. Conway & Huffcutt, 1997; Scullen et al., 2003).

Rater Group by Dimension and General Performance Factor

When rater groups and dimensions are reviewed together, the findings are equally confusing. Some studies show that the highest order distinction raters make is between technical and non-technical dimensions, and these generalize across rater groups and rating instruments (e.g. Facteau & Craig, 2001; Scullen et al., 2003). Since subordinates are interested in a complex assortment of leadership behaviors including interpersonal relations and professional knowledge when assessing their leaders (Fox & Bizman, 1988), and are more concerned with the bosses' ability and willingness to help them develop technical and administrative competencies than the leader's overall performance (Scullen et al., 2003). It is not surprising, then, to find that subordinate ratings are highly correlated across task and contextual dimensions (Scullen, et al., 2003) and that peers and subordinates provide comparable ratings for the contextual behaviors, such as peers and subordinates agreeing on a leader's team building skills (Maurer, et al., 1998).

Alternatively, research disputes generalizability of rater groups and/or dimensions. It is found that more knowledgeable raters more highly differentiate dimensions (Borman, 1987), peer and supervisor ratings are equivalent at the dimension level (Viswesvaran et al., 2005) and dimensions account for most of the variance in MSF ratings (e.g., Mount et al., 1998; Scullen et

al., 2000). Disputing that, however, is research that peers most highly emphasize interpersonal facilitation when assessing effective leadership (e.g. Conway, 1999; Fox & Bizman, 1988; Hooijberg & Choi, 2000) and most strongly distinguish between the task and contextual behaviors among rater groups. Conversely, bosses more strongly emphasize objective measures (task dimension) such as reaching goals or staying within budget constraints (e.g., Conway, 1999; Johnson & Meade, 2010; Scullen et al., 2003). Therefore, supervisors show a stronger and more consistent relationship between ratings of the task dimension and perceived leadership effectiveness in their employees, but not between the contextual dimension and perceived leadership effectiveness (Johnson & Meade, 2010).

Finally, the general factor or higher order dimension of leadership effectiveness/performance, offers its own set of contrasting findings. Research finds rater groups interpret the leadership effectiveness dimension similarly and it is generalizable (Woehr et al., 2005), but other studies find that different sources have somewhat different perspectives on overall performance (e.g. Scullen et al., 2003; Conway & Huffcutt, 1997).

In reviewing all of the disparities, it is apparent that rater group comparisons at the dimension and general factor levels are fraught with controversy. From a research and practical perspective, the prevalent use of MSF for leadership assessment in U.S. organizations requires a rigorous review of the data, methods, results and application of results.

Despite decades of extolling the value of MSF for performance assessment (since Lawler, 1967) and the intuitive belief that it is a good idea, it is clear that many of the assumptions of MSF are untested (Church & Bracken, 1997). Some researchers claim there is little value contributed to the process through examining rater group differences (Viswesvaran et al., 2005),

others argue that rater source differences exist and merit further research (e.g. Gregarus et al., 2003; Hoffman et al., 2010). In sum, if ILTs exist, and create significant group differences in ratings to support the value in administering and using MSF as the most prevalent leadership assessment tool in the U.S., then why is the research not consistent?

Prominent methodologists propose that the results of MSF leadership research to date are obscured due to inappropriate analytic methods (Hoffman, et al., 2010). Along with the MSF factor and rater group model controversies presented above, the other principal flaw in group comparison research is the failure to appropriately test the rating instruments for measurement invariance. If the rating instrument is not equivalent across all sources, then any interpretations as to the extent that raters agree or disagree may be inaccurate or misleading (e.g. Vandenberg & Lance, 2000; Woehr et al., 2005). The next section describes the MI testing process and the value it brings to supporting the comparison of rater groups to identifying where and why group differences might exist.

Measurement Equivalence and Invariance

To make any statements as to the similarities or differences between rater groups, it is imperative that we test for measurement equivalence/invariance (ME/I) . In sum, ME/I is a confirmation that the measures we use are measuring the same attribute under different conditions (Dierdorff, Surface, Meade, Thompson, & Martin, 2006). In the case of leadership MSF, it is confirmation that the organization groups from which we are gathering our data are assessing the same number and types of attributes (dimensions) of the person being rated and that each group's cognitive frames of reference are interpreting the items and measures in the survey instrument in the same way.

Establishing full measurement equivalence between groups and scales involves ascertaining both conceptual and psychometric equivalence (Cheung, 1999). Conceptual ME tests that the same number of dimensions are being used across rater groups to similarly assess behaviors, and that these assessments are of the same weight or magnitude between groups (Diefendorff, Silverman, & Gregarus, 2005). This is done by establishing configural and psychometric (metric) invariance. Configural invariance posits that the items in the measures load on the appropriate dimensions across rating sources, and metric invariance posits that the relation between specific items and the underlying dimension is the same across rater groups. Both of these conditions are necessary to establish conceptual equivalence (Woehr et al., 2005).

Psychometric ME tests affirm that the instrument items are responded to in the same manner by different rater sources. This includes the same degree of reliability, variance, range, mean levels of ratings and intercorrelations across rater sources for items and scales and dimensions (Diefendorf et al., 2005). These tests are not necessary for establishing ME, but they do confirm the same amount of variance across rater groups, establishing that the sources are

equally reliable (Woehr et al., 2005). Inequivalence with psychometric tests may indicate where rating biases are present (Deifendorf et al., 2005).

From a practical perspective, establishing configural MI is critical. It ensures the organization that the results from the MSF survey are based on valid group differences, not measurement error resulting from different interpretations of the instrument by different rater groups. As an example, if the raters are assessing the interpersonal behaviors of the leader and there are five scales assessing these behaviors, then the raters must cognitively agree that these five scales adequately represent relating behaviors, and that each of the scales carries the same magnitude of importance in their minds. If raters do not agree on the number of factors or the magnitude of these factors, ME recognizes these discrepancies among group interpretations and consequently, methodologically sound group comparisons cannot be made. If the initial tests for MI do establish configural invariance, then additional tests for metric equivalence will add valuable information about the variances and patterns of responses between groups. This helps reveal why differences exist among rater groups (true mean differences and/or group biases) and supports the final research conclusions. As discussed earlier, simply assessing the means between groups is not sufficient because the patterns of responses between groups might carry significant meanings even if the mean differences do not. Therefore, a full range of ME testing, before applying the MSF results in developmental or administrative decisions, is critical to organizations in order to ensure the group results are appropriate for true comparison, and not the result of inequivalent measures.

Although a large coalition of eminent scholars support establishing conceptual and psychometric ME before reporting and comparing group results (e.g. Cheung, 1999; Vandenberg & Lance, 2000; Woehr, et al., 2005), this testing is not a prominent part of the MSF literature. In

the past, it was difficult to test for invariance with the statistical tools available. This is no longer the case. Confirmatory Factor Analysis (CFA) is a superior method of establishing MI (Cheung, 1999), particularly when testing large samples with multiple scales and dimensions and relationships between factors (Meade & Lautenschlager, 2004). More recently, scholars have found that the original omnibus testing procedure for establishing MI, which was considered the only test necessary if the data were found invariant, is not valid with large sample sizes. This is ironic as large sample sizes are a requirement for MI testing, but also raises questions as to the results of any previous leadership MSF research that assessed and validated MI via the omnibus test.

The contemporary approach to assessing MI involves a series of CFA tests, using the omnibus test as simply a baseline model. From there, the MI tests compare a set of parameters in a more constrained or "nested" model (parameters are set to be equal) with a less constrained model (parameters are allowed to differ), and then compare the resulting model fit indices to see if the fit is reduced in the more constrained model. If model fit does not worsen in the constrained model, the samples are invariant (Dierdorff et al., 2006). This is true except for the test for scalar invariance where model fit should worsen (Vandenberg, 2010).

The key to establishing MI comes from assessing configural invariance (Horn & McArdle, 1992). This test compares nested models to ensure an equal number of latent factors across samples (same factor structure) and that the same items load on the same factors across samples by examining the patterns of fixed and free factor loadings between groups. If the same factor pattern matrix appears between groups and results in strong fit indices, then there is reasonable empirical support for equivalent cognitive frames of reference and support for

configural invariance (Vandenberg, 2002). If configural invariance is not upheld, the measures are judged to be non-equivalent and no group comparisons can be made (Dierdorff et al., 2006).

If configural invariance is present, then a test of metric invariance (Horn & McArdle, 1992) examines the equivalence of items' factor loadings across groups, i.e., do the items relate to the latent factors in the same way across groups. This test examines differences in magnitudes of factor loadings of like items between groups. If both constraining and freely estimating the loadings of like items that are thought to be equal across groups results in equally strong model fit with the same factor variances, then these are accepted as reasonable empirical representations of the conceptual scaling used by the raters when responding and reasonable tests of the equality of the conceptual frames of reference. Therefore, metric invariance is supported.

If configural and metric invariance are upheld then conceptual equivalence is established. The perspective on further testing for additional psychometric equivalence is mixed, but MSF leadership research is interested in specific between group differences. Therefore, a full range of ME testing of the error variances associated with the rating items helps assess that raters are rating the same dimension across sources, that the loadings of the rating from the rating sources are of equal magnitude and that the variances are equal. Non-equivalence in these additional psychometric ME/I tests do not indicate non-equivalence across sources but rather provide more detailed information about where and why group differences might exist (Diefendorff et al, 2005).

In continuing the steps, next is a test for scalar invariance to assess the intercepts of like items' regressions on the latent variables. If theoretically we expect groups to differ, we want to constrain the intercepts and find poor fit and move on to latent variable/structural invariance tests. Current research also uses the scalar invariance test as a way to examine the means

between items (Vandenberg, 2010). In most research this test is infrequently conducted and not often needed based on the substantive context of the study (Vandenberg & Lance, 2000), but a significant part of MSF research at this time focuses on "if" group differences exists in leadership assessment, and so the test for scalar invariance (no differences between groups) is appropriate to help establish this point. From here, the researcher chooses other tests of invariance to support the theory and findings. These include tests for: factor variances (constrain factor variances as equal across rater groups), factor covariances (constrain factor dimensions as equal across groups) and factor means (constrain factor means as equal and unequal across groups) (Diefendorff, et al., 2005).

A summary of the recommended list and sequence of CFA invariance tests (R. J. Vandenberg, personal communication, May 17, 2010) and the related null hypothesis and meaning for each as they pertain to MSF ratings (Diefendorff et al., 2005) is below:

0. *Omnibus:* tests equality of covariance matrices across groups. The null hypothesis is that the covariance matrices are invariant. Rejecting the null hypothesis requires further testing to find where the non-equivalence exists. Failure to reject the null indicates overall measurement equivalence across rater groups and it was previously thought that no further testing was required. Recent findings, however, show that this test is sensitive to large sample sizes and results should be interpreted and used cautiously (R. J. Vandenberg, personal communication, May 17, 2010).
1. *Configural invariance:* tests for equivalent factor structure across groups. The null hypothesis is that the fixed and free factor loading patterns are the same from one group to the next. Rejection of the null hypothesis implies that the respondent groups disagree

over the number or composition of factors in the instrument based on differences in implicit theories, differences in access to performance information, and/or disagreement about the duties of the person they are rating, and that no group comparisons are valid because the underlying constructs are different between rater groups. Failure to reject the null indicates that the rater groups are using the same frame of reference and might be compared with regards to latent mean group differences, and that further tests should be conducted.

2. *Metric Invariance*: tests for equality of scaling units across groups. The null hypothesis that strength of association of an item to the true score is invariant across groups, that is, the relationship of specific behaviors/items to a certain dimension is the same for different rater sources. Rejecting the null indicates that items load differently on factors across rater groups. This test is very sensitive and somewhat controversial, and often will not support invariance. Further testing should be run by removing items with lower factor loadings (R.J. Vandenberg, personal communication, May 17, 2010).
3. *Scalar invariance*: tests for invariance of item intercepts across groups. Tests null hypothesis that intercepts on like items' regression on the latent variables are invariant across groups, that is, that the means of item indicators are equivalent across rater groups. Rejecting the null might indicate that sources agree conceptually on the dimensions of the instrument, but one source consistently rates higher or lower. This could be a difference in groups, or rater bias. Scalar invariance testing is now also a way to test the means (Vandenberg, 2010).
4. *Factor Variances*: tests for equality of latent construct variances across groups. Tests null hypothesis that factor variances are invariant across groups, i.e. the groups are using

equivalent ranges of the construct continuum. If the null is rejected, then groups with smaller factor variances are using a narrower range on the construct continuum. If this fits well, then report these means.

5. *Equal and Unequal Factor Means*: two tests set for establishing equal and unequal latent factor means across groups. Null hypothesis is invariant factor means across groups. Rejection of the null indicates that the rater sources are rating the employee at different levels on the latent construct. This test evaluates the presence of lenient or severe rater sources at the construct level and/or group mean differences in ratings.

In summary, establishing configural MI is vital for any study comparing group differences. Employing the full range of contemporary ME/I tests provides a comprehensive way to evaluate ratings equivalence across rater sources and dimensions of interest, as well as support rater group comparison results. However, the MSF literature stream spans over 30 years and the concept of testing for measurement invariance in group comparison research has been in the methods literature for over 20 years, yet MI is "poorly understood" (Meade & Bauer, 2007, p. 612) and inadequately applied (Vandenberg & Lance, 2000). Unfortunately, with the equivalence of measures simply "assumed" away by researchers, any group comparison research is suspect (Vandenberg, 2002, p. 140).

The complexity of employing MI testing is no longer a valid excuse for bypassing this critical step, and therefore begs the question of why MI testing is still absent from the literature. One reason posited by Vandenberg (2002) for this oversight is that researchers are concerned that their data will not support configural invariance, thereby preventing comparison of the groups for similarities and/or differences. If the researcher is unable to support configural

invariance, then the inability to make group comparisons due to inequivalence renders the data collection and research study moot.

Another challenge is that proper ME/I testing requires large sample sizes, n>600, (Vandenberg & Self, 1993). There are currently very few between group field samples in leadership assessment that are large enough for testing (Epitropaki & Martin, 2004). There is one large database that dominates the MSF leadership field, however, and includes rater group sample sizes sufficient for MI testing. The database is described as:

"A widely used 360 feedback instrument that has been extensively researched in the literature. It is used primarily for leadership development purposes…(and) is based on research on how successful managers learn, grow and change. It has been subjected to a number of validation studies and has received favorable reviews as a valid measure of leadership behavior."(Atwater et al., 2009, p. 879)

A review of the literature failed to turn up any validation of this database outside of its owners, The Center for Creative Leadership (CCL), and internal researchers (e.g. Leslie & Fleenor, 1998; Spangler, 2003). The only instance found in the seminal MSF leadership literature where this database had undergone MI testing was Scullen et al. (2003), where the data passed the omnibus test and the measures were deemed equivalent. In addition, while the purpose of this is survey is described in the literature as primarily developmental, the same items and factors are used to assess leadership performance, i.e., for administrative purposes. The literature makes a strong distinction between the application of these two types of feedback (e.g. Borman, 1997; Scullen et al., 2003), yet more than 90% of companies that use MSF surveys apply the results to a variety of critical human resource management areas. That, coupled with the concern that the primary MSF leadership database is primarily composed of white ($n \geq 87\%$) males ($n \geq 68\%$) (Scullen et al., 2003) who are no longer representative of the demographics of the U.S. workforce (Cascio & Aguinis, 2008) provides an excellent opportunity for initiating a body of

MSF leadership research where strong methodological processes are applied to a new and robust set of 360 degree data.

The next section chronologically highlights the existing seminal MSF leadership empirical literature, many of which use the older database outlined above. The overview highlights the samples, models and methods in each study, and the discrepancies in results from one study to the next, even when using the same data.

Detailed MSF Leadership Assessment Studies

MSF leadership research has reached its current contradictory state through three general areas of discrepant findings. They include the raters themselves (both the individual characteristics of the raters and the rater groups), the items and dimensions of leadership skills, behaviors and characteristics (including sub-factors and general factors) and methodological challenges such as measurement error and/or lack of MI testing.

Given that my research focuses on rater groups and performance factors, the following section chronologically outlines the seminal empirical MSF pieces in these areas. Each study includes a summary of the underlying theory and purpose of each study, the dimensions and factors of interest, the sample, the group comparison results and the comparison of results to previous research.

Mount et al. (1998)

This study extended an earlier taxonomy of leadership roles proposed by Mann (1965) that included three broad factors and sixteen sub-dimensions:

1. Administrative Role: Planning, organizing, personal organization, and time management
2. Human Relations Role: Conflict Management, Listening, Motivating Others, Human Relations, Personal Adaptability
3. Technical Skills and Motivation Role: Problem analysis, Oral Communication, Written Communication, Personal Motivation, Financial & Quantitative, Occupational and Technical Knowledge

Mount et al.'s study included 2,350 managers (self) and 16,450 raters (two peers, two subordinates, and two bosses), who were primarily white (87%) and male (74%). Each person

completed a management skills assessment instrument used for management development consulting. The purpose of the study was to compare results between and within rater groups by using more than one rater in each category. The best fitting model for testing consisted of the three leadership roles and the seven raters, equaling ten factors.

Their main conclusion was that individual raters were more responsible for variance in the findings than rater groups, precluding aggregation of ratings within rater groups, except for the supervisor's group. They proposed that bosses used a more common frame of reference than the other rater groups, based on their experience and training, and therefore these results could be aggregated due to the similarity of within group ratings. The study did not use any form of testing for measurement invariance.

Conway (1999)

This meta analysis extended Borman and Motowidlo's (1993) task (job specific behaviors) and contextual (non-specific job behaviors) dimensions. The purpose was to establish the value and contribution of contextual performance over and above task performance. Conway identified 14 MSF managerial performance studies and proposed job dedication and interpersonal facilitation as dimensions of a contextual construct, and technical-administrative task performance and leadership task performance as dimensions of a task construct. The job dedication component included commitment, ethics, effort, and confronting problems and the interpersonal component included relationships, compassion, sensitivity, cooperation and consideration. The technical and administrative task component included forecasting, budgeting, planning, hiring, knowledge, decision-making and resourcefulness, and the leadership task component included motivation, supervision, human relations and power. The study also looked

at the contextual and task dimensions as part of a general factor of overall performance. Conway used self, peer and supervisor rater groups.

The results showed a distinct difference between the contextual and task constructs of leadership by rating source, with peers making the strongest distinction. Moreover, rating sources placed more importance on the dimensions when assessing overall performance. For example, peers paid more attention to contextual performance (interpersonal facilitation and dedication) and supervisors paid more attention to the sub-dimension of leadership task performance. This was attributed to peers having more opportunity to observe manager interpersonal skills, managers making more of an effort to showcase these skills in the presence of their peers, and supervisors relying more on results than observed behaviors. Self-raters did not make a distinction between the two constructs. Self-raters did, however, place emphasis on the technical-administrative dimension of overall performance, but there was almost no correlation with the interpersonal dimension. As a meta-analysis, the study did not address any MI.

Scullen et al. (2003)

This study extended the Mann (1965) three skill model (technical, human, and administrative), the Borman and Motowidlo (1993) two dimension model (job specific-task and non-job specific contextual behaviors) and Conway's (1999) contextual (interpersonal facilitation and job dedication) dimension of performance. The purpose was threefold: to determine the construct validity of a model hybrid consisting of four measures of manager performance (technical skills, administrative skills, human skills and citizenship behaviors); to examine the interrelationships of the four measures in an attempt to establish higher order

dimensions of task and contextual behaviors; and to assess the generalizability of the findings across rater groups.

Scullen et al. used data from two multi-rater feedback instruments that had been used "extensively" in practice (Scullen et al., 2003, p. 54) as performance feedback tools, and used confirmatory factor analysis (CFA) to conceptually aggregate the items and dimensions to represent the four constructs of interest. They confirmed the four factors and the two higher order factors of task and contextual performance. They also more specifically identified task performance as the ability to transform raw materials and support and maintain the technical core of the organization, and included both technical skills (specialized knowledge, skills and abilities and quantitative managerial functions) and administrative skills (think and act in organization system, including people, structures, procedures, and policies in order to achieve objectives). The contextual performance construct was defined as supporting broader organizational, social and psychological environments in which the technical core must function. The two factors in this dimension were human skills (ability to work with and through people to accomplish goals) and citizenship behaviors (three types of extra role behaviors: interpersonal, organizational, and job task conscientiousness).

The Scullen et al. (2003) study compared the two data sets of managers, with a minimum N=3,424 in one, and a minimum N=1546 in the other. Most of the samples included leaders with more than two rater responses per organization group, but the researchers only used data from a maximum of two random raters in each organization group. Rater demographics were predominantly white (87; 90%) and male (68; 74%). Results found the four lower order constructs (technical skills, administrative skills, human skills, citizenship behaviors) were distinct but generalizable across the four rater perspectives (boss, peer, subordinate, self). This

finding disputed an earlier supposition that different rater groups held different definitions and conceptions of performance (Borman, 1997), and supported another study proposing that rater groups shared a common conceptualization of performance (Facteau & Craig, 2001). Scullen et al. (2003) also suggested the four lower order constructs lent themselves to the two higher order dimensions of task and contextual performance, but the two higher order dimensions did not lend themselves to predicting an even higher global leadership effectiveness factor. They proposed, therefore, that the highest distinction raters make is between technical and interpersonal performance. They also proposed that individual rater effects were very strong (much like Mount et al., 1998), yet outside the scope of their study. They performed the recommended (at the time) omnibus test for the equivalence of covariance matrices across groups (Vandenberg & Lance, 2000) and found support for invariance across measures in each instrument. When testing the measures using only one rater per group, however, the authors found that in each instrument the factor variances and covariances were not equivalent across all rater perspectives. Further examination showed that except for technical skills, self-ratings were consistently lower than other ratings, indicating that self-raters use a smaller continuum to rate performance in all categories except technical skills.

Viswesvaran et al. (2005)

This meta analysis examines MSF research to address the "disagreement in the literature as to whether there is a general factor in job performance ratings" (p. 108) after controlling for measurement error. Theoretically, they proposed that there was a general factor, mostly comprised of contextual behaviors. The study also addressed previous empirical studies on rater level effects/group differences, and the idea that different rater levels observed and emphasized different aspects of performance, thereby causing their ratings to address different constructs.

Viswesvaran et al. (2005) cited the Maurer, Raju & Collins (1998) and Facteau & Craig (2001) studies using confirmatory factor analysis (CFA) and item response theory (IRT) to establish measurement invariance between rater groups, which led them to surmise there was no rater group effect and the same dimensions underlie peer and supervisor data and correlate at nearly 1.0. As such, Viswesvaran et al. (2005) supported between group supervisor and peer rater aggregation to replace missing scores within supervisor group in their meta analysis.

Viswesvaran et al.'s (2005) results also found that peers and supervisors had a substantial amount of measurement error in their general factor ratings (63% and 33%, respectively), primarily due to "halo effect". Halo effect is a psychological process wherein a rater forms an opinion of the person they are rating, and that opinion is unique to that rater, i.e., some of the opinion will overlap other raters and some will not, creating measurement error. This supported Scullen, Mount & Goff's (2000) findings that peer ratings included more halo effect than supervisor ratings. After correcting for halo error and other measurement error, however, the general factor of performance accounted for 60% of the true score variance. Also of importance was finding that group factors accounted for 40% of the variance, although group factor variance was not part of the study's purpose. Additionally, they proposed that all theories need to incorporate a general factor to be of merit, and that sub-dimensions might be aggregated into an overall general performance factor. This directly contradicted Scullen et al., (2003) regarding the distinction of sub-dimensions and a general performance factor among rater groups. Later research disputed the findings based on the specified model, which did not use rater groups as factors (Hoffman et al., 2010).

Woehr, Sheehan, and Bennett (2005)

This study examines measurement equivalence across military rater sources by comparing the fit of a model of MSF ratings as a function of rating source factor and underlying general performance dimensions. While the study evaluated ME across sources (self, peer and supervisor) the total research sample was only N=1028, with only one rater from each group. Woehr et al. (2005) found conceptual equivalence among sources for the performance constructs but did not find error variance equivalence. Results showed that dimension effects were bigger than source effects, but unique variance was higher than either performance or source effects. This study is included in this present research because it was the only one found that tested for MI past the omnibus test, although the sample size in each rater group was far below the N>600 threshold. Some other things to note from this study are that the data was collected for research (not administrative) purposes, each dimension was assessed with only one item (contributing to the potential for significant measurement error) and raters had received frame of reference training, which has been shown to effect ratings (Moses, Hollenbeck, & Sorcher, 1993). Additionally, the performance dimension factors were comprised of military job performance measures, including military appearance.

Hoffman et al. (2010)

This study extends the work of Conway (1999), and Borman and Brush (1993). The purpose was to further distinguish idiosyncratic rater from rater source effects. The study employed a hybrid taxonomy of managerial performance, derived from previously mentioned models, as well as one dimension they added, which they felt worked better with their data. The three broad performance dimensions were leadership and supervision (leading employees,

confronting employee problems, participative management, and change management), technical activities and mechanics of management (resourcefulness, being a quick study, and decisiveness) and interpersonal dealings and communication (compassion and sensitivity, building and mending relationships, straightforwardness and composure, self awareness, putting people at ease, differences matter).

The study used two management/leadership development samples, each of which was used in other studies outlined above. One consisted of 22,240 managers and 156,940 raters, primarily white (76%) and male (64%). The other was 2,350 managers and 16,450 raters (same sample used by Mount et al., 1998), primarily white (87%) and male (74%). The four rater groups included 2 supervisors, 2 peers, 2 followers and self.

Hoffman et al. (2010) tested six different factor structure models to see how individual raters, rater groups and performance dimensions contributed to the latent structure of MSF ratings. Their findings supported variances across: raters, rater source, dimensions and a general performance dimension, respectively. The magnitude of the variances by factor differed from previous studies in two key ways. One is that rater source accounted for a much higher source of variance than found in previous studies (and three times that of performance dimensions). The second is that the higher order dimension of general performance accounted for a much lower portion of variance (4%) compared to previous studies such as Viswesvaran et al. (2005), who found 27% in their meta analysis, and Scullen et al. (2000), who found 14%. The discrepancy between the former was attributed to Viswesvaran et al.'s lack of inclusion of rater source effects in their model, whereas Hoffman et al. hypothesized large source effects. The discrepancy between Hoffman et al. (2010) and the latter was attributed to Scullen et al.'s (2000) use of

parameterization, which causes the misattribution of source-specific variance to trait factors. This study did not address MI testing.

As highlighted by the seminal studies above, there are significant differences in MSF research results as to the impact of rater groups and dimensions of interest. Scullen et al. (2003) found that rater group results were the lowest source of variance, Viswesvaran et al. (2005) did not include rater source effects in the performance model, and Hoffman et al. (2010) found that rating sources provided the second highest source of variance. In addition, Viswesvaran et al. (2005) promoted aggregating results even between rater source group, yet Hoffman et al. (2010) discouraged aggregating rater source results even within groups.

The models and dimensions of interest in each study varied greatly, too. Mount et al. (1998) used a leadership role taxonomy (Mann, 1965) with three factors and sixteen sub-dimensions, Conway (1999) used a two factor model with three sub-dimensions and added an extra role behavior sub-dimension, Scullen et al. (2003) used a hybrid of Mann (1965) and Conway (1999) for the sub-dimensions and Borman and Motowidlo (1993) for the two higher order factors. Viswesvaran et al. (2005) used an older Viswesvaran, Ones, & Schmidt (1996) job performance framework with nine dimensions, and Hoffman et al. (2010) used a hybrid model with three dimensions, two from previous studies and one they assigned based on their internal review of the data.

The overall lack of agreement on the models used in leadership MSF research is cause for great concern since the misspecification of models is the most important source of error for researchers when assessing their findings (Cheung & Rensvold, 2001), particularly in group comparison research (Hoffman, et al., 2010). Each empirical study above recognizes that MSF

ratings are a function of the raters and the dimensions, but each model offers a completely different set of factors for testing. Additionally, these are primarily post hoc based on the data set. Thus, the inconsistency of leadership MSF findings with respect to impact, amount of impact and convergence or non-convergence of ratings across groups is not surprising.

This present research, however, puts forth a leadership assessment model that addresses both the rater group and dimensions of interest. First, this research models five rating sources, something not currently found in the literature although it is used in practice. Second, the dimensions in the model incorporate what Conway (1999) referred to as the non-leadership oriented tasks specific to managerial work and explicit goal achievement and the non-specific behaviors of leadership like showing personal concern and building relationships. Since Mann (1965), relationship behaviors and skill-based task performance have consistently emerged in the leadership group comparison literature. Other dimensions have been put forth and modeled with no consistent results (i.e., Conway, 1999 *Dedication*; Viswesvaran, et al., 2005, *Effort*), but relationship behaviors and skill-based task performances have unfailingly provided a framework for categorizing and assessing a vast majority of leadership behaviors and skills, regardless of the data. Additionally, an exploratory factor analysis of the survey items will help determine the presence of any other latent factors and a separate leadership effectiveness dimension will be tested to provide a comprehensive model of MSF leadership based on the rater groups and consistent dimensions of interest.

Along with model misspecification, the lack of appropriate measurement equivalence testing in most MSF studies greatly contributes to the confusion surrounding rater group comparisons. If the rating instrument is not equivalent across the sources then any interpretations of the ratings and any practical application of the results may be inaccurate and misleading

because there is no support for the level of agreement or the consistency of construct interpretation across sources (Woehr et al., 2005). The present research completes a full range of ME tests (configural, metric and error variance) on the data collection instrument to assure conceptual and psychometric equivalence before proceeding with any group comparisons.

Without a consistent framework of dimensions for evaluation and the inclusion of all organization rater groups as part of the model, it is impossible to determine where group comparisons exist in MSF leadership ratings and the impact of each. Moreover, without appropriate and consistent ME testing of the instrument, there is the even more important question of if group comparisons can be made. The research questions and hypotheses for this current study follow.

Research Questions and Hypotheses

Multisource ratings typically expose differences among raters (Cheung, 1999). Before those differences can be properly identified, understood and applied, however, it is critical to take the appropriate steps to ensure measurement equivalence. Measures invoking cognitive processes, like MSF, establish *a priori* reasons for conducting ME/I procedures, and demonstrating invariance across groups is a logical prerequisite to conducting cross-group comparisons (Vandenberg & Lance, 2002). It is apparent from the extant literature that this critical step is often not applied, or applied incorrectly based on the current state of the literature. In fact, previous research findings suggest that the data will not support invariance between all groups (Lance & Bennett, 1997) and factors (e.g. Epitropaki & Martin, 2004; Scullen et al., 2003). This failure to test for MI renders any substantive comparisons of group ratings inappropriate (Woehr et al., 2005).

My initial set of hypotheses focus on an overall assessment of ME between groups. I hypothesize there will be configural invariance between all groups and dimensions with a null hypothesis:

Hypothesis$_{(0)}$: The model of leadership performance will be equivalent across self, follower, peer, supervisor and supervisor's boss rater groups.

and the alternative hypothesis:

Hypothesis$_{(a)}$: The model of leadership performance will be non-equivalent across self, follower, peer, supervisor and supervisor's boss rater groups.

Establishing configural invariance by not rejecting the null hypothesis allows the testing of group comparison hypotheses. My first hypothesis deals with the value of collecting data from a rating group higher than supervisor. Earlier studies have supported using feedback from one rater group as a substitute for another rater group (Viswesvaran et al., 2005), and it is possible that supervisors and their bosses view the self group similarly. This, combined with the expense of administering MSF surveys would make for a valuable discovery.

Hypothesis 1: The supervisor and supervisor's boss group ratings for the relationship-based dimension, skill-based dimension and overall performance factor will be conceptually equivalent and highly correlated, and support between group aggregation of scores.

For the next set of hypotheses I also expect non-equivalent error variance relationships (MI tests 3-6) between leaders and other groups, indicating rater group biases, differences in means, range restriction, etc. This expectation is based on previous research that some groups are subject to defining and understanding performance dimensions differently based on their position in the company (Lord et al., 2001) and different opportunities to observe leaders' behaviors and skills (Woehr et al., 2005). Error variance inequivalence is not the same as configural inequivalence, and therefore does not preclude group comparisons, but it does indicate the presence of biases versus ILTs when comparing group differences. The results of these tests should address some of the inconsistencies between rater groups and dimensions found in the literature.

As an example, there is mixed information regarding self-evaluations in the MSF literature, and the MI tests for scalar invariance (mean differences) will address those discrepancies. Generally, individuals are not good at evaluating themselves (Harris & Schaubroeck, 1998) and lack insight into their behavior impact (vanDierendonck et al., 2007).

Alternately, self- raters are reported to be a reliable ratings' group (Barr & Raju, 2003) and self-evaluation of ability may correspond closely to performance (Mabe & West, 1982). Perhaps Conway (1999) said it best, "Self-ratings were somewhat puzzling (p.11)."

Hypothesis 2: When comparing leaders' self- ratings to direct reports, error variance inequivalence will be supported.

Hypothesis 3: When comparing leaders' self- ratings to peers, error variance inequivalence will be supported.

Hypothesis 4: When comparing leaders' self- ratings to supervisors, error variance inequivalence will not be supported.

In other group comparisons (direct report, peer, supervisor) Tsui and Ohlott (1988) found that overall leader effectiveness perception is not very different between peers, subordinates and superiors. More recently, Scullen et al. (2003) found that raters do not distinguish a higher order general performance/leadership effectiveness factor, and Johnson and Meade (2010) discovered a consistent relationship between supervisor's ratings of the task dimension and overall perceived leadership effectiveness in their employees, but not between the contextual dimension and overall perceived leadership effectiveness. This might be the most interesting set of hypotheses because in addition to the results listed above:

- Peers paid more attention to interpersonal facilitation when making overall leader performance ratings (Conway, 1999)
- Supervisors pay more attention to task performance when making overall leader performance ratings (e.g. Johnson & Meade, 2010; Conway, 1999)
- Different sources have somewhat different perspectives on performance (e.g. Scullen et al., 2003; Conway & Huffcutt, 1997)

Hypothesis 5: When comparing supervisor's ratings to direct reports, error variance inequivalence will be supported.

Hypothesis 6: When comparing supervisor's ratings to peers, error variance inequivalence will be supported.

Hypothesis 7: When comparing peer ratings to direct reports, error variance equivalence will be supported.

A consistent research finding shows that direct reports and peers pay more attention to interpersonal (contextual) leader behaviors and supervisors are more interested in skill-based behaviors. Therefore, I am testing a series of hypotheses about the correlations between the lower order dimensions and leadership effectiveness, by rater group.

Hypothesis 8: For peer ratings the relationship based dimension will be more strongly correlated to the overall leadership effectiveness factor than will the skill based dimension.

Hypothesis 9: For direct report ratings the relationship based dimension will be more strongly correlated to the overall leadership effectiveness factor than will the skill based dimension.

Hypothesis 10: For supervisor ratings the skill based dimension will be more strongly correlated to the overall leadership effectiveness factor than will the relationship based dimension.

Methods

Instruments

The Leadership Circle™ (TLC) is a leadership development consulting firm offering multisource feedback assessment and follow up development and coaching plans. The company consults with organizations, including government and military, worldwide to help them evaluate and enhance their leadership personnel. The TLC measure compares a selected employee's self-evaluation of her or his leadership skills and behaviors to the evaluations of their supervisor(s), supervisor's boss, peers and followers/direct reports.

It is common to use existing industry measures, as shown in the outlines of previous studies, so this type of empirical research is well represented in applied and theoretical leadership literature (Hoffman et al., 2010). The popularity of using applied data stems from the fact that leadership development ratings tend to be more accurate than administrative ratings (Scullen, et al., 2003), and collecting industry data yields much higher sample sizes for testing and analysis.

TLC's instrument is based on applied research and business consulting experience, and is strongly tied to three streams of theory – leadership, individual psychology and adult development. This is consistent with previously mentioned performance measures of leadership, which are psychological in nature (Viswesvaran et al., 2005), and the premise that most MSF leadership feedback is for developmental purposes.

Many of the reported MSF leadership research studies customize their data, even though they are using the same database from previous studies. This customization includes deleting items and changing the dimensions of interest, creating a lack of consistency in the leadership development frameworks used for each analysis. This present research is using dimensions with

historical theoretical and empirical support for their value in assessing leadership behaviors - relationship behaviors and skill-based task performance. Each of the previous studies offered one or more behavioral dimension, labeled interpersonal, human relations, contextual, human skills, etc., and one or more skill-based dimensions, labeled productivity, technical tasks, administrative tasks, technical activities, etc. A specific review of TLC dimensions that address those areas is below, followed by my proposed categorization of the measures into my testing framework.

The total TLC data set includes 119 items, grouped into 29 scales and then eight summary competency and/or behavior dimensions. To adhere with the existing theoretical sub-dimensions of leadership effectiveness, relationship behaviors and skill-based task performance, I will be using items from four of the eight TLC summary dimensions, which measure leadership behaviors and achievement. The four dimensions have eleven scales and 43 items total. The dimensions, scales, number of items per scale and an example item from each scale follow.

Relating (5 scales, 18 items). Examines the leader's capability to relate to others in a way that brings out the best in people, groups and organizations.

1. Caring Connection (3 items): I am compassionate.
2. Fosters Team Play (3 items): I share leadership.
3. Collaborator (3 items): I work to find common ground.
4. Mentoring and Developing (4 items): I am a people builder/developer.
5. Interpersonal Intelligence (5 items): I display a high degree of skill in resolving conflict.

Authenticity (2 scales, 6 items). Examines the leader's capability to relate to others in an authentic, courageous and high integrity manner:

1. Integrity (3 items): I hold to my values during good times and bad.
2. Courageous Authenticity (3 items): I speak directly even on controversial issues.

Systems awareness (1 scale, 3 items). Examines the leader's focus on whole system improvement and productivity:

1. Systems Thinker (3 items): I reduce activities that waste resources.

Achieving (3 scales, 16 items): Examines the leader's ability to offer visionary, authentic, and high achievement leadership.

1. Strategic Focus (9 items): I focus in quickly on the key issues
2. Achieves Results (4 items): I pursue results with drive and energy.
3. Decisiveness (3 items): I make decisions in a timely manner.

I am also using TLC's five item global measure for overall leadership effectiveness. The items from this measure are:

1. I am satisfied with the quality of leadership that I provide.
2. I am the kind of leader that others should aspire to become.
3. I am an example of an ideal leader.
4. My leadership helps this organization to thrive.
5. Overall, I provide very effective leadership.

Each scale in TLC's instrument uses a 5-point Likert scale response (*1 = Never, 2 = Seldom, 3 = Sometimes, 4 = Often* and *5 = Always*). The survey also gathers demographic information including industry (health care, consulting, manufacturing, etc.), culture, language, gender, age, education level, ethnicity, management level, and profession (medical professional,

operations/production). The survey is administered electronically to all participants and requires answers for all items.

Participants

The participants are working managers identified and selected by their organizations for participation in a comprehensive leadership development program. This program includes collecting MSF ratings from five sources, consulting with the leader to discuss the results, and developing coaching plans when warranted. Raters are chosen by the leader in training, but specific responses are anonymous except for those from the supervisor and the supervisor's boss.

The database consists of 6,557 English speaking leaders (self). The leaders choose raters from four rater groups, including supervisor's boss, boss, peers, and followers. The total "others" sample (supervisor's boss, boss, peer, direct report) consists of 61,986 raters, yielding a total of 68,543 leaders and raters. The breakdown by rater group is:

Rater Group	*N*
Supervisor's Boss	2,998
Boss	7,364
Follower/Direct Report	22,839
Peer	28,785
Self (leader)	6,557
Total N	68,543

These are English-speaking leaders from U.S. based businesses. They are 42.9% female, with a mean age of 44.8 years and predominantly White, 87.4%. Over 90% of the self group has a college education level above an associate's degree, and these leaders represent a diverse cross section of industries.

Procedures

Before beginning my data analyses I explored my database using SPSS 18.0 to check for missing data and outliers and normality, perform a series of exploratory factor analyses to determine which items to retain for subsequent analyses, assess scale reliabilities, and review correlations for dimension and rater group aggregation decisions.

To review the data I ran descriptives and made missing data replacement or deletion decisions. My choices for making missing data decisions were using a mean rating replacement derived from all raters across that perspective (Scullen et al., 2003), deleting the data pairwise for a missing variable (the least dispersion around my true scores) (Roth & Switzer, 1995) or deleting listwise which eliminates the entire case if there is any missing data (Field, 2011). Additionally, I also looked for outliers in the data set. My final data decisions regarding replacement or deletion of data were made based on the amount and sources of the problematic data. I expect minimal missing data problems because the survey is administered electronically to participants with an interest in the results.

I used exploratory factor analysis (EFA) to evaluate how the items from the database load on my proposed theoretical dimensions – relationship behaviors and skill-based task performance. This was done to address the adequacy of the underlying items in the dimensions and the appropriateness of the dimensions for assessing leadership performance (Hurley, Scandura, Schriesheim, Brannick, Seers, Vandenberg, & Williams, 1997). My items have not been subjected to any previous statistical analyses so I am approaching them differently than the previous MSF studies cited in this study. A review of the best practices methodological literature explains that exploratory techniques help to describe, summarize and/or reduce the initial sample (Hurley et al., 1997). If a sample has not been adequately and objectively described it is difficult

to support the use of the results in the absence of MI testing. Since the overarching research question in this study addresses the measurement invariance of the instrument (and it is a question that needs to be addressed in all group comparison research), it is imperative that the items and factors are appropriately tested before undergoing the MI process since they have not been independently used or validated in any previous research. Earlier cited studies reference the deletion of items and changing of factors, but the researchers start with CFA's which are only appropriate when the underlying theory and measurement model are well-defined (Hurley et al., 1997). Although the broad theory behind MSF ratings (number of dimensions raters observe) might hold in these cases, the disagreements about what these dimensions of interest are and the frequent deletion of items from the original data sets would seem to strongly indicate that EFA is the proper starting point (e.g. Field, 2011; Hurley et al., 1997). Therefore, both EFA and CFA are necessary in this study for assistance in exploring the items, assigning the appropriate dimensions and supporting the final results.

My research compares responses between groups after confirming measurement equivalence, so while the psychometric properties of the instruments are less relevant (Scullen et al., 2003), it is helpful to support my study with some straightforward content validity. To do this, I will supplement the EFA with an informal, two-part Q-sort process (Stephenson, 1935) using Subject Matter Experts (SME's). This step allows me to replicate using SME's, a process many previous MSF studies use for dimension assignment, and compare my EFA statistical results with my SME results.

I will also use Cronbach's alpha, a widely accepted statistical estimate of the correlations between two random samples of items in a measure, to determine the internal consistency of my resulting dimensions. While a reliable measure showing substantial internal consistency does not

necessarily provide a complete test of validity, it does provide a baseline for beginning interpretation (Cronbach, 1951). A generally accepted Cronbach's alpha level to support reliability is >.70 for new measures (Cortina, 1993), so I expect to find higher levels for my scales. If not, I will review the individual items and consider dropping those whose deletion improves the measure, and that do not fit my theoretical constructs. Dropping items based on sound theoretical and methodological reasoning is a well-accepted practice in MSF research when using application-based data (Hoffman et al., 2010). If I do exclude any TLC data, I will re-examine the reliabilities of my measures and report my decisions and processes (Williams & O'Boyle, 2008).

I also want to assess my within and between group correlations so I can address one ongoing debate in the literature regarding aggregation. The MSF literature frequently investigates aggregation to review within and between group agreement and/or for data replacement needs. Aggregating within groups uses a mean score for the group (i.e., all peer level scores) versus using each individual score in the group (Johnson & Meade, 2010). Interrater Agreement (IRA) correlations measure the consensus, or variability, between raters within a group to determine if a mean group score is representative of the group as a whole, or if individual raters within the group differ to an extent that scores should be evaluated separately. For my IRA calculation I will use $r_{(wg),}$ which compares actual to expected variance (James, Demaree, & Wolf, 1984). High scores, $r_{(wg)} > .70$, indicate within group agreement and help support within group aggregation of scores. However, for the importance of the decisions based on MSF results, a level of IRA >.90 is more appropriate (LeBreton and Senter, 2008). I will also assess the Pearson correlations (relative consistency) between rater groups, known as Interrater Reliability (IRR), to address the second aggregation dispute in the literature – aggregating

between groups or using one rater group's results to replace a missing rater group's results. To support my Pearson correlations, I will use ANOVA to calculate the Intraclass Correlations (ICCs), or proportion of variance in the ratings due to the between group differences. Specifically, I will calculate *ICC(1,K)* in order to index the agreement among my raters (J.M. LeBreton, personal communication, January 26, 2011). *ICC(1)* evaluates if group membership effects ratings and *ICC(K)* evaluates how reliably the mean ratings distinguish rater groups. High ICC values, (ICC > .80) confirm both consensus (IRA) and consistency (IRR) among raters and targets (LeBreton & Senter, 2008). The results of all the correlation calculations above will address the literature discrepancies about within and between group aggregation and address Hypothesis 1 regarding between group aggregation.

For my MI data analysis I will use M*plus v.6* (Muthen & Muthen, 2010) to do a series of seven CFA's using Structural Equation Modeling (SEM). Researchers typically use SEM to explore the links between the latent variables (unobserved variables) and measured variables (indicators), and investigate the relations between the latent variables (Burnette & Williams, 2005). This SEM output provides a comprehensive analytical overview of the model because it combines factor analysis and regression simultaneously (Williams, Vandenberg, & Edwards, 2009). In the case of this research, though, I will be investigating between model changes to establish if and where MI exists between groups,

I will begin with a baseline model and null hypothesis that each source is equivalent, i.e. the covariance matrices are invariant. This is the previously recommended first test (omnibus) in the ME series (Vandenberg & Lance, 2000) and constrains everything to be equal across all groups. Failure to reject the null indicates equivalence across all groups and in the past was used to support measurement equivalence, unequivocally. Research over the past decade, however,

finds that the omnibus test is sensitive to large sample sizes and so it is no longer the quintessential MI test, but it does provide a basis for comparison among other models (R. J. Vandenberg, personal communication, May 17, 2010).

After the omnibus I will continue my MI testing and address my error equivalence Hypotheses 2-7. To support making between group comparisons, I must establish configural ME by testing for agreement among rater groups regarding the number and/or composition of factors. If disagreements do not exist, I will test for metric invariance to look at disagreements over the pattern of factor loadings across rater groups. In other words, do raters agree on the relative importance of the indicators in defining that particular dimension/factor. If both configural and metric invariance are upheld, then conceptual equivalence is established. This will address my equivalence hypotheses and support my ANOVA and correlation group data results. Due to the nature of MSF research for application purposes it is important to continue the ME tests to help understand where and why differences exist between rater groups. Results from the next three ME tests address where differences lie in rater group responses due to bias, the amount of error variance for dimensions across groups and the scale range across groups. First, the test for scalar invariance examines the different mean levels across groups. This test will be of particular interest in helping to understand where differences exist in MSF between rater groups. I expect to find mean level differences between the self group and the direct report, peer and supervisor groups (Hypotheses 2-4) based on existing research results. Next, a test for invariant factor variances reviews if the rater groups use the same continuum range of scale for responses and/or violate homogeneity of variance, again offering support for Hypotheses 2-4. Finally, MI tests for equal and unequal factor means to assess the impact of lenient or severe rater biases and look for true mean differences between rater groups and dimension ratings. The results of these tests will

address Hypotheses 5-7 (self-ratings). The last set of hypotheses (7-9) test for differences between rater groups based on the relationship and skill dimensions and general leadership effectiveness factor, and the full series of MI test results, along with my preliminary data analyses, will address those.

The MI testing process assesses model fit in one test to model fit in a more restricted next test. Typically in a CFA a series of industry-accepted goodness of fit statistics evaluate the adequacy of the model in question based on the cut off value statistics below.

Rule of thumb cut off values (Hu & Bentler, 1999):

1. **Chi Square statistic (X^2):** Chi square statistics provide a test of model fit for the baseline model, and the chi-square value directly relates to the sample size (Bentler & Bonett, 1980). Large samples and different sample sizes are susceptible to errors in this test, so I will combine my chi square findings with other fit comparison indices to most effectively detect model differences (Cheung & Rensvold, 1999).
2. **Comparative Fit Index - CFI (Bentler, 1990) and Tucker-Lewis Index - TLI (Tucker & Lewis, 1973)** – relative fit indices that evaluate model fit relative to a null model and take into account the overall number of model parameters estimated. Cutoff: $\geq$ .90
3. **Root Mean Square Error of Approximation - RMSEA (Steiger, 1990)** – lack of fit per degree of freedom. Cutoff: $\leq$ .08
4. **Standardized Root Mean Squared Residual -** SRMSR – summary index of the percentage of variance unaccounted for by the fitted model. Cutoff: $\leq$ .1

In the case of contemporary Measurement Invariance CFA testing, however, only a subset of the Hu and Bentler (1999) statistics are used, X^2 and CFI for model goodness of fit and

ΔX^2 and ΔCFI for between model comparisons, because the point of interest in MI is the change in model fit throughout the testing procedure (Vandenberg, 2010). For my particular research I will use an even smaller set of statistics, CFI and ΔCFI, because I have large and uneven sample sizes which do not work well in X^2 statistical formulas when comparing models. Therefore I will be assessing if CFI≥.90, indicating good model fit, and if ΔCFI < .01, indicating little change in between model fit (Cheung & Rensvold, 1999).

The results from my MI tests and preliminary data analyses are described in the next section.

Results

After exploring the data I removed cases with government, military and education industry identification. Previous literature supports that these three industries have unique perspectives on leadership (Woehr et. al, 2005) and therefore had the potential to obfuscate any generalizability of the industry results. This left a sample size of N=48,137.

	Frequency	Percent
Self	4477	9.3
Supervisor	4979	10.3
Direct Report	16,726	34.7
Peer	19,986	41.5
Supervisor's Boss	1969	4.1
Total N	48,137	100.0

A visual review of the P_P plot (probability-probability plot) graphs and simple histograms for the items, by user group, revealed no outliers values. The P-P plot graphs plot the cumulative probability of a variable against the cumulative probability of a normal distribution (Field, 2011). For each variable the data points fall very close to the diagonal, or the normal distribution line. The histograms plot a single variable value against the frequency of the scores (Field, 2011) and also showed normal looking distributions. Table 1 a-e reports the descriptive by rater group and item, including mean, standard deviation, missing data, and extreme high and low results. One item in the mentoring and developing measure -*I help direct reports create development plans* (REMen450)- has a high percentage of missing data (6%-24%) for all rater groups compared to any of the other items. It is the only item in the data set targeting a leader's behavior towards a specific rater group, which helps explain the amount of missing responses since this behavior might not be observable by the other rater groups. I chose to remove that from the dataset. Additional data screening for the remaining items shows missing data is most

prevalent in items like mentoring and strategy, which are also not necessarily observable by all rater groups (i.e., supervisor's boss and direct reports, respectively), but do not target a specific group. Another interesting observation when exploring the data is that only the self-rater group has high extreme values. Since no other items specifically mention another rater group, no other items have missing data > 12% and a review of the variances for each item (by rater group) and visual graphs show no extreme values, I left the rest of the items in the dataset for the EFA process.

EFA

My database has not been used in previous research so the preliminary data analyses included extensive EFA testing. First, I chose a random sample of 500 self raters and let the 42 items load with no factor, rotation or item restrictions. A review of the eigenvalues (seven factors > 1), variance explained, cross loadings and scree test (showing two or three factors to be retained) led me to testing a series of three factor and two factor models with all varieties of rotation and suppression. All EFA's yielded a Kaiser-Meyer-Olkin (KMO) > .95, indicating compact patterns of correlations and the adequacy of my samples in yielding distinct and reliable factors (Field, 2011). Additionally, Bartlett's measure, testing the null hypothesis that the correlation matrix is an identity matrix, was significant at $p \leq .05$, confirming that there are relationships between the variables. The only cross loadings were the five items in the leadership effectiveness measure. I tested the reliability of those items as a measure, and found $\alpha > .95$. The EFA and scale reliability results confirmed my expectation that these five items did not assess individual attributes and behaviors but rather outcomes from these attributes and behaviors soI removed them from the EFA testing and treated them as a separate factor. The results of the leadership effectiveness testing also lends support to previous research that leadership sub-

dimensions are strongly correlated to a higher leadership effectiveness/performance factor, andthe strong cross loadings and high scale reliability made theoretical sense.

After reviewing my sample EFA results I decided to run EFA's across the complete database and all rater groups. After reviewing the results from a variety of condition and rotation constraints, my optimal solution of choice for all rater groups (considering eigenvalues, variance explained, factor loadings, scree tests and rotations) was a three factor model with oblimin rotation. Table 2 contains a summary of the results of the EFA's.

The EFA factors for factor 1 and factor 3 interchanged between the self group and the other rater groups. In self, factor 1 is a decisive-based dimension, including items surveying decision- making, directiveness and courageous skills and behaviors. Factor 3 is an analytical/strategic-based dimension including strategy, systems thinking, trend-spotting, anticipation, results and integration items. For the other rater groups, however, factor 1 is the analytical/strategic dimension, including strategy and systems thinking. Factor 3 includes some of items that loaded in factor 1 for self, directiveness and courageousness, but the three decision-making items that loaded for self did not load consistently across any of the other rater groups. Factor 2, however, is the same latent factor and includes exactly the same items between all five rater groups. It is an interpersonal-based dimension, measuring positivity, negotiating, compassion, caring, teamwork, growth and development and shared leadership. Since measurement invariance testing is the basis for this research, I needed a set of items that loaded consistently across all rater groups. After a final EFA review, that was 23 out of the original 43items.

Since MSF is grounded in differences between rater groups, and the concepts of measurement invariance and implicit theories are based on the subjective, cognitive differences

in perceptions between rater groups, using a Q sort process provided qualitative support to my EFA results. To begin my Q sort I enlisted 12 SME's, gave them a card for each TLC item that loaded for three or more rater groups (30 items) and asked them to group the items based on similarities. Seven SME's responded, and despite extensive literature proposing that raters really only view four or less dimensions (e.g. Conway, 1999; Scullen et.al, 2003), these raters found between three and ten dimensions, with a mode of four and a mean of five. Another observation was that the key words for different items (i.e., strategy) did not always influence the sorting process. As an example, four items included the word strategy, and only two raters put those four items in the same dimension. On the other hand, raters did show a consistency in grouping items with the words "courageous" (3 items), "feeling" (3 items), "employee development" (2 items), and "results" (2 items). These 10 items were items that loaded across all rater groups in the EFA as well, and three of these four sub-factors contained behavioral rather than skill-based items. There was also some additional consistency in the placement of other interpersonal items and also some analytical skills..

From these Q sort results and my EFA analysis, I developed two additional Q sort pages. One listed the 30 items and the dimension definitions I assigned based on the results from my first Q sort SME test (Courageous, Feeling, Development, Results). The second list used the dimensions I assigned to my three factors from the EFA results – analytical, interpersonal and courageous. Then I asked another group of 15 SME's to place the items in which category they felt each belonged, in the two separate tests. Ten responded, and I cross-referenced my Q sort results with my EFA factors and item results. The Q sort process supported the content validity of the EFA results since items that loaded strongly across all rater groups did the same hereand the weaker loading items from the EFA results were not as consistently placed in the Q sorts.

These results of the Q sort tests also supported the EFA dimension assignments for the analytical, interpersonal and courageous dimensions.

To summarize the data exploration results, I began with 48 items. One of those referenced a specific rater group and was missing a relatively large number of responses compared to any other items. Since this item was not directly observable by all rater groups and missing a relatively large number of responses I deleted it. Next, my EFA revealed significant cross loading for the five leadership effectiveness items, regardless of the rotation or rater group, indicating these items were strongly rooted in all of the underlying latent factors. To support these five items as a separate measure I assessed Cronbach's Alpha for measure's reliability on its own and it yielded $\alpha > .95$ so I established a separate leadership effectiveness dimension. The EFA results for the remaining items (42) yielded a good number of items that did not load across rater groups. I selected the items that did, 23, along with those that loaded across three or more rater groups (30 items total) and subjected them to a two phase SME Q sort. The first phase yielded anywhere from 3-10 factors, with a mode of four factors and a mean of five. From these results I conducted another Q sort with another group of SME's. This one asked participants to group the 30 items under the four factors from the first Q sort and the 30 items and under the three factors from the EFA. These results consistently confirmed my EFA results for items that loaded consistently across all five rater groups and for items that did not load consistently across all five rater groups. The EFA results are presented in Tables 2a-e.

Reliability of items and dimensions

With the final 23 items and three dimensions and my leadership effectiveness items and dimension, I conducted scale reliability tests to determine Cronbach's α for each dimension and each rater group. Table 3 lists the reliability of each dimension by each rater group, and none

of the scales had improved reliability by deleting any of the items. When testing the reliability of scales, it is an industry standard to expect α >.8 for established scales and α >.7 for new scales (Cortina, 1993). The rigor of the EFA and Q-sort processes proved valuable as the Cronbach's α's were high for each dimension, and for each rater group: analytic dimension $\alpha \geq .85$, interpersonal dimension $\alpha \geq .88$, courageous dimension $\alpha \geq .79$ and leadership effectiveness dimension $\alpha \geq .87$.

Correlations

Based on the statistical analysis so far it appears that aggregating my items within the dimensions is appropriate. To provide additional support for the α values in the preceding section, however, I also reviewed the number of items in each dimension and the relationships between the items in each dimension by rater group. Table 4 reports the results. Since all items are significantly correlated within dimensions (p's < .01), the dimensions had high scale reliabilities with no item deletions and the EFA and Q sort processes revealed three factors and the leadership effectiveness dimension I ran descriptives with the reduced item set and dimensions of interest. Table 5 reports a summary of the cases by dimension and by rater group.

Table 6 reports the Pearson correlations between dimensions and rater groups. All dimension correlations were significant in each rater group, and part of this study is specifically aimed at the correlations between the dimensions of interest and leadership effectiveness. For the self group there was a significant relationship between the analytical dimension and leadership effectiveness dimension, $r = .60$ ($p < .01$), interpersonal and leadership effectiveness dimension, $r = .60$ ($p < .01$) and courageous and leadership effectiveness dimension, $r = .49$ ($p < .01$). These correlations are much lower than in all the other groups, just as the self scale reliabilities were much lower than any of the other groups.

In the supervisor group there was a significant relationship between the analytical dimension and leadership effectiveness dimension, $r = .77$ ($p < .01$), interpersonal and leadership effectiveness dimension, $r = .82$ ($p < .01$) and courageous and leadership effectiveness dimension, $r = .69$ ($p < .01$). For the direct report group there was a significant relationship between the analytical dimension and leadership effectiveness dimension, $r = .84$ ($p < .01$), interpersonal and leadership effectiveness dimension, $r = .87$ ($p < .01$) and courageous and leadership effectiveness dimension, $r = .71$ ($p < .01$). The overall correlation results in the direct report group were highest when comparing dimensions and leadership effectiveness versus the other groups. In the peer group there was a significant relationship between the analytical dimension and leadership effectiveness dimension, $r = .82$ ($p < .01$), interpersonal and leadership effectiveness dimension, $r = .85$ ($p < .01$) and courageous and leadership effectiveness dimension, $r = .68$ ($p < .01$). Finally, in the supervisor's boss rater group there was a significant relationship between the analytical dimension and leadership effectiveness dimension, $r = .79$ ($p < .01$), interpersonal and leadership effectiveness dimension, $r = .84$ ($p < .01$) and courageous and leadership effectiveness dimension, $r = .70$ ($p < .01$). In each rater group, the interpersonal dimension was the most highly correlated with leadership effectiveness.

ANOVA by dimensions and rater groups

The theory of MSF is rooted in between group comparisons and a significant amount of research in leadership MSF addresses the validity of within and between group aggregation, so correctly analyzing and comparing the means of each group is critical. To compare the group means I ran a series of ANOVA's by dimension and rater group. ANOVA tests assume normality within groups and homogeneity of variance between groups. While my descriptives did indicate normality, comparing uneven group sample sizes tends to violate the homogeneity of

variance assumption and the standard F-test (tests to see if the standard deviations between two populations are equal) ANOVA statistic is not designed to handle this violation (Field, 2011). Therefore, I used ANOVA tests designed to account for violation of the homogeneity of variance. First, I ran Levene's test. This tests the null hypothesis that the variances of the groups are the same. A significant result ($p < .05$) indicates the variances are significantly different. In my ANOVA Levene's test was significant for all dimensions so my groups violate the assumption of homogeneity of variance. If the data violates the assumption of homogeneity but is normal, statistical best practices suggest using Welch's *F* statistic (Field, 2011). Welch's *F* was significant among all four dimensions at:

Analytical $F(4, 7871.93)$, $p < .001$

Interpersonal $F(4, 8052.7)$, $p < .001$

Courageous $F(4, 7868.83)$, $p < .001$

Leadership Effectiveness $F(4, 7871.93)$, $p < .001$

I used *post hoc* tests to compare means among all different combinations of the groups while controlling for error. There are a variety of *post hoc* tests available, each with its own strengths and weaknesses. The most powerful is the Games-Howell procedure and it is best suited for large and uneven sample sizes like mine (Field, 2011). The results of Games-Howell, summarized below and detailed in Table 7, yielded significant and non-significant mean differences between dimensions and rater groups. For the analytical dimension:

Self (0) was significantly different from all other groups, $p < .01$

Boss (1) was significantly different than self, direct reports and peers, $p < .01$

Direct report (2) was significantly different than all other groups, $p < .01$

Peer (3) was significantly different than all other groups, $p < .01$

Supervisor's Boss (5) was significantly different than self, direct report and peers, $p < .01$

In summary for the analytical dimension, the self, direct report, and peer groups all had significant mean differences from all the other rater groups, and the supervisor and supervisor's boss groups did not have significant mean differences. In the interpersonal dimension:

Self (0) was significantly different than peer and supervisor's boss, $p < .01$

Direct report (2) was significantly different than peer and supervisor's boss, $p < .01$

Peer (3) was significantly different than self and direct report , $p < .01$

Supervisor's boss (5) was significantly different than self and direct reports, $p < .01$

Boss (1) was not significantly different than any other groups, $p > .01$

In summary for the interpersonal dimension, both self and direct report were significantly different from peer and supervisor's boss but not significantly different from each other, and supervisor was not significantly different from any other rater group. For the courageous dimension:

Self (0) was significantly different than direct report and supervisor's boss, $p < .01$

Boss (1) was significantly different than direct report and supervisor's boss, $p > .01$

Direct report (2) was significantly different than all other groups, $p < .01$

Peer (3) was significantly different than direct reports and supervisor's boss, $p < .01$

Supervisor's boss (5) was significantly different than all other groups, $p < .01$

In summary for the courageous dimension, self, boss and peer were all significantly different from direct report and supervisor's boss, and in turn those two groups were significantly different than all of the other groups. In the leadership effectiveness dimension all rater groups had significantly different means except peer and supervisor's boss.

Intraclass Correlation Coefficients (ICC's)

To explore the effect of group membership, I also calculated r_{wg}'s and ICC's (1,k). The r_{wg} statistics will add additional information to the aggregation decisions within groups and determine the extent to which random rater response or group leniency bias play a role in MSF ratings. ICC(1) estimates the interrater agreement (IRA) and interrater reliability (IRR) of individual raters, otherwise known as the quality of the individual ratings. ICC(k) estimates the stability of the ratings, that is the consensus and consistency (IRR + IRA), of the rater group (J.M. LeBreton, personal communication, October 28, 2011).

R_{wg} is a component of IRA and IRR, and adds to the credibility of the ICC's and pinpoints specific within group information. Here I am specifically testing for random rater response and leniency bias tendencies within groups. An example of random rater response might be found if the supervisor's boss asked the supervisor to complete the survey for her/him, and/or if a rater became bored or was not interested in participating and just randomly completed the survey. Reviewing the data to this point it did not appear likely that there would be this problem, but it is statistically easy to confirm. To do this I compared each item's observed variance to the expected variance (observed variance/2) in a uniform null distribution if random

rater response was present. Using the arguable industry accepted minimum threshold of $r_{wg} > .7$ (LeBreton & Senter, 2008), each dimension and rater group had more than 89% of their responses exceed the rule of thumb, $r_{wg} > .7$.

Leniency bias is the tendency for raters not to judge somebody they know as objectively as they should based on rater group membership. An example would be a direct report rating a supervisor who was responsible for promotion and retention decisions. This r_{wg} is calculated using a slightly skewed distribution, as would be expected if leniency bias was present. While a large percentage of raters in each rater group, 74-95%, had an $r_{wg} > .7$, the two groups exhibiting the largest number of raters with leniency bias were self and direct report. The results for the random rater and leniency bias r_{wg} results are in Table 8.

Table 9 reports single and average measure ICC's. As expected, the single measure ICC(1)'s are lower than the average measures ICC(k)'s, but both the single and average measures are significant in each dimension and for each rater group at $p < .05$. The single measures, ICC(1)'s are useful for determining the quality of the individual ratings in each group and providing estimates of the effect size, or extent to which group membership influenced the ratings. Benchmarks for evaluating effect sizes come from Cohen (1988) and are:

r = .10, small effect size, explains 1% of total variance

r = .30, medium effect size, explains 9% of total variance

r = .50, large effect size, explains 25% of the variance

In all dimensions, self had the lowest effect size by group:

Analytical ICC(1) = .38
Interpersonal ICC(1) = .37
Courageous ICC(1) = .49
Leadership Effectiveness ICC(1) = .46

Additional results for single measures, all but one which indicated high effect sizes, included:

Supervisor:
Analytical ICC(1) = .55
Interpersonal ICC(1) = .57
Courageous ICC(1) = .54
Leadership Effectiveness ICC(1) = .77

Direct Report:
Analytical ICC(1) = .56
Interpersonal ICC(1) = .61
Courageous ICC(1) = .48
Leadership Effectiveness ICC(1) = .83

Peer:
Analytical ICC(1) = .57
Interpersonal ICC(1) = .62
Courageous ICC(1) = .53
Leadership Effectiveness ICC(1) = .79

Supervisor's Boss:
Analytical ICC(1) = .57
Interpersonal ICC(1) = .60
Courageous ICC(1) = .55
Leadership Effectiveness ICC(1) = .79

For the average measures, ICC(K)'s were also very high, indicating high levels of IRR and IRA, and once again supporting stability of the mean ratings in the group (J.M. LeBreton, personal communication, October 23, 2011).

Self:
Analytical ICC(K) = .83
Interpersonal ICC(K) = .87
Courageous ICC(K) = .79
Leadership Effectiveness ICC(K) = .81

Supervisor:
Analytical ICC(K) = .91
Interpersonal ICC(K) = .94
Courageous ICC(K) = .82
Leadership Effectiveness ICC(K) = .95

Direct Report:

Analytical ICC(K) = .91
Interpersonal ICC(K) = .95
Courageous ICC(K) = .79
Leadership Effectiveness ICC(K) = .96

Peer:

Analytical ICC(K) = .92
Interpersonal ICC(K) = .94
Courageous ICC(K) = .82
Leadership Effectiveness ICC(K) = .95

Supervisor's Boss:

Analytical ICC(K) = .91
Interpersonal ICC(K) = .94
Courageous ICC(K) = .83
Leadership Effectiveness ICC(K) = .95

Measurement Invariance Tests

After full exploration of the data and reporting preliminary results (results similar to what are found in many other MSF leadership research publications), I ran my MI tests using a baseline model of all five rater groups and four dimensions of interest to see if MI supported my earlier findings for comparison purposes. Table 11 below lists each test, the results and the conclusion from the results for this databases. Due to the instability of χ^2 in large and uneven sample sizes and pursuant to Vandenberg (2010) only CFI and ΔCFI statistics are pertinent to this series of MI tests. The benchmarks for the conclusions are CFI > .90 and ΔCFI < .01.

Table 11: Measurement Invariance Tests and Results			
Model	CFI	ΔCFI	Conclusion
Model 0: Omnibus* Constrain items variances and factor covariances to be equal across groups Test for equal variances and covariances	.82	-	Not expected to find MI due to large sample size and χ^2
Model 1: Configural Invariance** Fix latent means to zero across groups and factors Test for equivalent factor structures between groups	.92	-	CFI>.90 Form of factor structure is equivalent across rater sources, may compare groups
Model 2: Metric Invariance** Constrain factor loadings to be equal across groups Test for strength of an item to the true score	.92	-	CFI>.90 Items to true scores are invariant across rater sources
Model 3: Scalar Invariance Constrain item intercepts to be zero Test for mean differences between latent variables	.11	-.81	CFI<.90, ΔCFI >.1 Mean differences exist somewhere between rater groups
Model 4: Unequal Factor Means No constraints Test for "very real" meaningful differences vs. Model 3	.67	+.56	CFI<.90, ΔCFI >.1 Mean differences supported as fit is improved over scalar
Model 5: Equal Factor Variances Factor variances are constrained to be equal across groups Test for the relationship (invariance) of the scales across groups and homogeneity of variance as well as range restriction	.54	-.13	CFI<.90, ΔCFI >.1 Homogeneity of variance is violated
Model 6: Equal Factor Means Mean factor ratings are constrained to be equal across groups Test for group differences in rating levels on the latent construct	.08	-.46	CFI<.90, ΔCFI >.1 Significant worsening in fit relative to Model 5, indicating real mean differences exist, possibly due to leniency or severity bias

*Test used as model baseline only, susceptible to error in large samples
**Configural + Metric Invariance = Conceptual Invariance

The first baseline test is the omnibus, or Model 0, which tests for equal variances and covariances between factors and rater groups. In earlier group comparison literature this test was the cornerstone of MI testing. If it produced good fit values the entire measure was declared equivalent, the MI testing would stop and the group comparison results were reported as valid. Recent MI research disputes the value of the omnibus test since it is highly susceptible to producing incorrect results with large sample sizes but it does provide a good MI baseline model (e.g. Cheung & Rensvold, 1999; Vandenberg, 2010). Since I do have a relatively large sample size Model 0 shows expectedly poor fit with CFI<.90.

Configural invariance testing is the new foundation for MI assessment and confirms that rater sources are in agreement about the number and/or composition of factors in the measurement tool. My result indicated acceptable model fit with five rater groups and three dimensions with CFI>.90. By establishing configural invariance with my model and data I can methodologically support the comparison of my group results since I cannot reject my null hypothesis:

Hypothesis$_{(0)}$: The model of leadership performance will be equivalent across self, follower, peer, supervisor and supervisor's boss rater groups.

The Model 2 test for metric invariance confirms the consistent strength of the relationships between items and dimensions across rater groups. For this data the metric invariance test resulted in acceptable model fit at CFI>.90. The metric invariance test tends to be sensitive, since it is item based, and will usually not support invariance without item removal. The initial data exploration and EFA process in this research removed problem items, however, so there is support for metric invariance. Establishing configural and metric invariance establishes the second important step in MI testing, conceptual equivalence, and allows me to address Hypothesis 1:

Hypothesis 1: The supervisor and supervisor's boss group ratings for the relationship-based dimension, skill-based dimension and overall performance factor will be conceptually equivalent and highly correlated, and support between group aggregation of scores.

This hypothesis is supported because there is conceptual equivalence and the correlation calculations, reported in Table 6, are nearly identical.

Correlations to Leadership Effectiveness	Analytical	Interpersonal
Supervisor	.77	.82
Supervisor's Boss	.79	.84

Further MI testing allows the researcher to investigate where differences lie between rater groups and why, i.e. tests for error variance equivalence. The scalar invariance test, Model 3, has been updated since the Vandenberg and Lance (2000) piece and is now a way to test the means. There are two different MI choices for the scalar invariance test, based on what the theory dictates. In this case, choice one (constrain the invariance, i.e. fix the latent means to "0") was used since theoretical differences should exist at the observed and latent variable levels in MSF research. Using the constrained model should yield an overall bad model fit and significant change in model fit from Model 2, which it does CFI=.11 and ΔCFI =-.81. These results indicate there are mean rater group differences, confirming the ANOVA results found in Table 7.

Continuing the MI tests, Model 4 (unequal factor means) should have improved fit from Model 3 (scalar invariance) if meaningful differences exist and it does, CFI=.54 and ΔCFI=.56. Model 5 investigates the homogeneity of variance and supports that the scale as a whole is invariant across groups. In this case CFI=.54 and ΔCFI=.-.13, indicating that the homogeneity of variance is violated for this set of rater groups. This finding confirms earlier ANOVA results and the choice of Games-Howell post hoc testing (reported in Table 7). Additionally, a poor fit for Model 5 indicates the presence of range restriction. Self is reported to rate themselves on smaller continuum than other groups and the large number of significant mean differences in Table 7 between self and others might contribute to the poor fit. Finally, Model 6 (equal factor means), tests whether or not mean differences exist, and if there are real differences the fit should worsen compared to Model 5, which it does, CFI=.08 and ΔCFI=-.46. The worsening fit usually dictates the presence of leniency and/or severity bias. Table 8 reports the severity/leniency r_{wg} test results for this data set, and the self and direct report groups had the highest percentages of raters exhibiting leniency bias in all three dimensions.

While configural and conceptual invariance were supported in tests of Models 1 and 2, the full complement of MI tests supported there is inequivalent error variance in the data. These results address my error variance Hypotheses 2-7.

Hypothesis 2: When comparing leaders' self- ratings to direct reports, error variance inequivalence will be supported.

Hypothesis 3: When comparing leaders' self- ratings to peers, error variance inequivalence will be supported.

Hypothesis 4: When comparing leaders' self- ratings to supervisors, error variance inequivalence will not be supported.

Hypothesis 5: When comparing supervisor's ratings to direct reports, error variance inequivalence will be supported.

Hypothesis 6: When comparing supervisor's ratings to peers, error variance inequivalence will be supported.

Hypothesis 7: When comparing peer ratings to direct reports, error variance equivalence will be supported.

The table below reports rater group and dimension means, which can also be found in Table 5, and Table 7 reports the statistically significant mean differences between groups and dimensions. For this research, Hypotheses 2 and 3 are supported since there is inequivalence in the error variance tests and self and direct report had significant mean differences in the analytical dimension (M= 3.76 and M=4.02) and in the courageous dimension (M=3.87 and M=3.96), respectively. Self and direct report were also the two groups with the highest occurrence of leniency bias, as shown by the r_{wg} results in Table 8.

Group Means	Analytical	Interpersonal	Courageous	Leadership Effectiveness
Self	3.76	3.99	3.87	3.68
Supervisor	3.85	3.96	3.88	3.99
Direct Report	4.02	3.98	3.96	4.12
Peer	3.88	3.93	3.87	3.96
Supervisor	3.82	3.92	3.80	3.92

Hypothesis 4 is partially supported. There is the presence of error variance in the MI testing, but in ANOVA testing results shown in Table 7 self was only significantly different than boss in the analytical dimension (M=3.76 and M=3.85). Hypothesis 5 is supported by the presence of error variance inequivalence and significant mean differences between boss and direct report for the analytical and courageous dimensions. Hypothesis 6 is only partially supported since peer is significantly different than boss only in the analytical dimension. Hypothesis 7 is not supported since equivalence was not supported in MI testing, and peer means were significantly lower than direct report means in all three dimensions (M_p=3.88, M_{dr}=4.02 – analytical; M_p=3.93, M_{dr}=3.98 – interpersonal; M_p=3.87, M=3.96$_{dr}$ – courageous).

Additionally, the establishment of configural invariance (Model 1) for the three dimensions and five rater groups and my correlation results in Table 6 allow me to address Hypotheses 8-10:

Hypothesis 8: For peer ratings the relationship based dimension will be more strongly correlated to the overall leadership effectiveness factor than will the skill based dimension.

Hypothesis 9: For direct report ratings the relationship based dimension will be more strongly correlated to the overall leadership effectiveness factor than will the skill based dimension.

Hypothesis 10: For supervisor ratings the skill based dimension will be more strongly correlated to the overall leadership effectiveness factor than will the relationship based dimension.

Hypotheses 8 and 9 were supported as the relationship (interpersonal) dimension was significantly more highly correlated than the skill based (analytical) dimension with the leadership effectiveness factor for both peer (r_p=.85 vs. r_p=.82, p<.01) and direct report rater groups (r_{dr}=.87 vs. r_{dr}=.84, p<.01). Surprisingly, as Table 6 shows, Hypothesis 10 was not supported. Previous research is unanimous on supervisor's interest in analytical skills over interpersonal skills when assessing overall leadership effectiveness. In this study, the

interpersonal dimension was more highly correlated with leadership effectiveness (r=.82 vs. r=.77, p<.01).

Discussion

This study supported conceptual measurement invariance between rater groups when analyzing three MSF leadership dimensions (analytical, interpersonal and courageous). However, further MI testing for error variance equivalence supported ANOVA findings of inconsistent between group mean differences, with only peers and direct reports showing significant mean differences in all three dimensions. In addition, EFA testing found that only 23 of the original 42 items loaded for all five rater groups.

My extensive preliminary data exploration bears significant discussion. The published MSF leadership literature does not include evidence of exploratory processes used before the CFA testing. The data item sorting and factor naming decisions are made by SME's in each study, even when the researchers are using data that has been previously published for the same purposes. It is clear from the research methods literature that this practice falls under the guise of exploratory factor analysis, not confirmatory factor analysis. In addition, there are several instances in the literature where items that have been used before are removed and/or assigned to new or different factors than in their previous research. While this is referred to as standard practice (Hoffman et al., 2010), it is strongly suggested that when items are removed from a dataset that an EFA is conducted to test the new factors (Field, 2011).

This exploratory process also yielded important rater group and dimension insight. In my study, three dimensions were strongly shared among the groups – analytical, interpersonal and courageous. Item loadings in these dimensions proved problematic, however, as the self group results were different than the other groups in two significant ways. First, the self rater group

items loading on factor one addressed courageous behaviors, which loaded on factor three for the other groups, and the items loading on factor three for self – analytical skills and behaviors – loaded on factor one for all other groups. An underlying premise in MI testing is that groups might not always interpret measures in a conceptually similar way and rating sources might not define performance in a similar way when rating on identical dimensions (Vandenberg & Lance, 2000), and in previous studies the self group often provided perplexing results (Conway, 1999) so these findings are similar to those reported elsewhere.

The EFA dimension results are also supported in the existing literature. The interpersonal dimension is the most consistent factor in MSF leadership research and it was the most consistent factor here, followed by a strong analytical dimension. While the factor name differs slightly from previous research - "task" to include planning, forecasting etc. (Conway, 1999) and "technical" to include knowledgeable, problem-solving/quantitative, etc. (e.g. Mann, 1965; Mount et al., 1998) - these items are similar to the skill and behavior items in my study that fell into the analytical dimension. These two dimensions are also exactly like those posited by Katz and Kahn (1976) who surmised that leaders needed analytical and strategic acumen and excellent interpersonal skills.

Initially it was surprising to find a third factor, courageous, but after further review, the items in this factor strongly resembled those in the dedication dimension found by Conway (1999). His dimension included items addressing commitment, ethics, effort, and confronting problems. The items in my courageous factor addressed directiveness and courage. Previous research had bundled the dedication items into a task performance dimension (Conway, 1999), but in both my case and Conway's these similar traits and behaviors comprised a strong, single factor. Once again, the preliminary data exploration in this study yielded important factor

information that might have been lost in a straightforward CFA procedure, and there is support in Conway's work (1999) that this factor is present in effective leadership assessments.

Finally, after assessing all the EFA results, 19 of the 42 items did not load across all rater groups. This is a substantial number of items to drop from the study, but since MSF is based on rater group differences it does not make sense to keep items that are obviously conceptually different between rater groups.

In sum, the EFA process yielded valuable information about items, factors, factor loadings and rater groups, and I believe using these results allowed my measures to pass the configural and metric invariance tests. I also believe my data are similar to what is found in other MSF leadership tools, and that calls into question the validity of those studies since none of them have used exploratory procedures.

The foundation of MSF is that that rater groups hold different ILT's across the dimensions of interest (Lord et al., 2001) and that assessing these differences provides a richer picture of the leader strengths and weaknesses. This study found that any significant mean differences depend on the groups and dimension of interest. Peers and direct reports were the only two groups with significant mean differences between them in all four factors, so in the true spirit of MSF for leadership development purposes, perhaps peers and direct reports are the only two groups whose comparisons would provide the most information about the leader when using measures that are invariant across rater sources.

There were also significant mean differences between the supervisor and direct report groups in all dimensions but interpersonal. Previous studies show that more importance is placed on interpersonal behaviors by direct reports than supervisors (Scullen et al., 2003), but that was not the case here. The importance placed on the interpersonal dimension by the supervisor

received additional support when the correlation between the interpersonal dimension and leadership effectiveness was higher than the correlation between the analytical dimension and leadership effectiveness dimension. In fact, all five rater groups had the highest correlations between the interpersonal dimension and the leadership effectiveness dimension.

Other research suggests peer ratings could be adequate substitutes for supervisor ratings (e.g. Lawler, 1967; Viswesvaran et al., 2005), but in comparing supervisor and peer groups there was a significant mean difference between them in the analytical and leadership effectiveness dimensions, but not in the interpersonal or courageous dimensions. Therefore, while peer ratings might work as proxy for supervisor ratings for the interpersonal and courageous factors, the mean differences in analytical and leadership effectiveness areas do not support this type of substitution.

Self ratings research takes up a lot of journal space and will always prove interesting to social science researchers. The self group in this study had the lowest correlations between dimensions, and the least amount of variance explained in the EFA results. Additionally, the self raters had significant mean differences from all other groups in the analytical factor, and significant mean differences when compared to the peer and supervisor's boss rater groups for the interpersonal factor. For the courageous dimension, self reported significant mean differences between the direct report and supervisor's boss groups, and in the leadership effectiveness dimension self had significant mean differences from all other groups. This array of statistically significant mean differences and the leniency bias in the r_{wg} calculations offer strong support as to why the MI tests did not pass error variance equivalence testing. Newer studies have proposed self is more accurate when asked to rate themselves the way their boss would rate them

(Schoorman & Mayer, 2008). Based on the inconsistency of self-other findings in this study, using the same instrument with the self group as with the other groups is not beneficial.

After reviewing the inconsistent statistical mean differences between rater groups and dimensions, it is also important to look at the practical mean differences. The highest mean difference between any two groups in any dimension was .44, and that was the self group low rating in leadership effectiveness, but the range for all other dimensions and groups was only .04 to .26. This calls into question the value of using MSF for leadership development purposes. The cost of purchasing, administering, analyzing and delivering the results requires the results to be of practical significance, not statistical significance. Very few organizations would be willing to invest so much into a process that yields, on average, less than a quarter of a point difference in means.

While this research goes a long way in uncovering the reasons why MSF leadership surveys continue to provide conflicting results, there are at least four noteworthy limitations I would like to address. First is in the naming of the leadership dimensions. My proposal outlined two broad dimensions, task-related and interpersonal-related, that I expected to find based on what currently existed in the literature. It was difficult at the time to name specifically these proposed dimensions because the field lacked consistency in identifying and adhering to a MSF leadership model. This study falls prey to the same problem. While the dimensions are very similar to most previously named dimensions and at least slightly similar to others, it was still difficult to use existing data and be able to fit the data to established names. This ongoing problem continues to contribute to the misspecification of models, one of the problematic parts of MSF leadership research (Hoffman et al., 2010), and it was obvious in this study why it persists.

There is an underlying consistency, though, in the raters' implicit perceptions of leadership behaviors/skills/competencies that also came through in my data. There is always a factor dealing with the relationship/interpersonal competencies of the leader, and always a factor addressing the analytical/technical skills and abilities of the leader. My research also produced a third factor, courageous. There is some support for this dimension via Conway's (1999) dedication dimension, but in the end I could not find existing factor names for any of the three dimensions that matched up to my items any more than the existing literature.

Also problematic to me during this research was the interchangeable use of the terms "leadership effectiveness," "leadership performance," "leadership development" and "leadership competencies," some even in the same study (Scullen et al., 2003). MSF makes a clear distinction between the use of ratings for administrative (performance) and developmental purposes. In a comprehensive review of the performance/development and management/non-management MSF literature, I could not find an instance where the terms were clearly defined or operationalized. Even within the same article, a self-described developmental MSF leadership tool recognizing interpersonal and task development opportunities became a performance evaluation system with a global dimension labeled leadership performance, and later in this same paper it became leadership effectiveness. Some researchers argue they are not the same thing, but at this point in the leadership MSF research history I am not sure you can successfully separate effectiveness and performance. From an ILT standpoint I do not think you can, either. So, from the murky perspective of what exactly defines and/or separates these two terms, my study had to rely on existing literature regarding both, even though I specifically examined portions of a leadership development tool whose focus and purpose is to assess overall leadership effectiveness.

Another limitation, or at least a legitimate challenge that could be lodged against this research, is the MI series used to test the data. To date there is no study incorporating the newly proposed MI procedures (Vandenberg, personal communication, May, 2010) and so the results and claims made here might not be popular. MSF studies that have tried to incorporate MI tests in the past have stopped at the omnibus test (Scullen, 2003). Further research in the field has found that the omnibus test is susceptible erroneous results in large sample sizes, yet correct MI tests require large sample sizes (Vandenberg & Lance, 2000).

Finally, my results are based on using extensive data exploration techniques, including EFA's, along with the complete series of MI tests. This is counter to the current state of the literature which engages only in the standard CFA processes. This creates a dilemma since some of my findings strongly support existing theory and results (i.e., interpersonal dimension being most highly correlated with leadership effectiveness in peer and direct report groups) and some of my findings strongly contradict existing research (i.e., analytical dimension being most highly correlated with leadership effectiveness in the supervisor rater group). The rigors of significant data exploration and MI testing methodologically support my results, but those same procedures also make them not necessarily comparable to published studies since those did not undergo the same processes. In essence, it proves that it is difficult to draw inferences and make claims about group differences as previously reported in the literature, even when using MI testing, if the rest of the field is not using the same methods.

Despite heading the list of most cited articles to appear in *Organizational Research Methods* (downloaded 9/7/2011 and again 1/12/2012), very few group related research projects run a full array of measurement invariance tests (Vandenberg & Lance, 2000), and even fewer are familiar with the new standards as put forth by Vandenberg (personal communication, May,

2010). In discussing this research with colleagues they always asked why this was the case. After conducting this research it seems there is both a research and application answer. The first is that after collecting or getting access to a large database of MSF leadership information, few researchers want to broach the idea that the data might violate MI and therefore not support group comparisons (Vandenberg, 2002). From an application perspective, few consulting companies or organizations engaged in the multi- billion dollar industry (purchasing and administering the instrument, analyzing the results, providing feedback and training, making and implementing critical human resource decisions) built on the premise that MSF is a valuable developmental and administrative decision making tool does not deliver the purported results. My findings support Vandenberg's concern. The preliminary data exploration pointed to problems with items loading across all factors and the perceptions of factors between rater groups, even before getting to the MI testing. The self group results alone, when compared to other groups regarding items, factors, low scale reliabilities, low dimension correlations and higher between group mean differences showcase why MSF studies continue to yield diverse findings.

These discrepancies also help spawn an abundance of research opportunities. There is clearly a need to develop a conceptual model framework for MSF leadership development. The interpersonal and analytical dimensions are consistently found in some manner across all studies, but only the interpersonal items loaded consistently between rater groups in this study. Further research could investigate the benefit of using measures of analytical ability across all groups when it appears that it is not as relevant to direct reports as it is to supervisors and supervisors' bosses. Along those same lines, it would prove interesting to test if the items that did not load across rater groups provided unique insight from that group. This would support separate

feedback tools for each rater group or a smaller subset of rater groups rather than a complete 360 survey, and perhaps even provide a more robust picture of the leader than what we think we currently get.

The most pressing need for future studies from a pure research perspective lies in proposing and testing a system for the appropriate treatment of MSF data, from data exploration to MI testing. The strongest story told by the data in this study came from the preliminary data exploration process, and the MI results provided support for the methodologically sound comparison of rater groups and the ANOVA results. If studies do not use MI testing the results cannot be held above suspicion or compared to other study results.

The most critical needs for future studies from a practical perspective is further investigation regarding how MSF tools are used in organizations. The correct tool must be used for the correct situation, developmental or administrative, and companies must ask the right questions about what value these assessments are ultimately bringing to the organization. Small mean differences between groups and dimensions of interest are not worth the expense of a widespread MSF leadership survey, but a comprehensive look at what items and dimensions register with which rater groups might yield an even more powerful instrument, or instruments, of measurement.

This research delivers three key advancements to the furthering of MSF in leadership development surveys. First, there is no prior leadership research addressing ILTs across all five of these organization groups: followers, peers, supervisor, supervisor's boss and self. It is posited that all organization levels vary in leadership prototype perceptions (Borman, 1997) but there is no empirical research addressing all five rater groups despite the use of supervisor boss ratings in some organizations, like universities. This study hypothesized that the supervisor's boss group

results would be similar to the supervisor group results, but that was not consistently true. Second, existing empirical rater assessments at the manager level used databases with a very large percentage of the raters and leaders being white males (e.g., Bartol et al., 2003; Scullen et al., 2003). There is an inherent bias in the appraisal system due to leadership prototypes invoking white male ILTs, which has caused disparate advancement for subgroups (Roberson, 2007). It is also shown that group similarity effects ratings (Rosette et al., 2008), and that rater groups might not respond to cognitive measures in the same way based on differing frames of reference such as gender (Vandenberg & Lance, 2000). As white males continue to decrease as a portion of the leadership and rater pool in a more heterogeneous workforce (Cascio & Aguinis, 2008), it is important to use data collected from a sample that is more representative of today's workforce. This study, using a database that is nearly 43% female, offers rater group comparisons taken from respondents who more closely model at least the professional female representation in the United States.

Finally, before making any results, inferences or generalizations of group differences or similarities in leadership assessments, the samples from each group must be large enough to test for ME (e.g. Epitropaki & Martin, 2004; Vandenberg, 2002). Each sample group in this study was large enough to provide adequate testing for measurement equivalence and therefore allow for methodologically sound group comparisons, something rarely found in the literature. As the U.S. continues to move from traditional hierarchical organizations to team and project based organization of work, the use of MSF will continue to increase (Viswesvaran et al., 2002). Addressing these three gaps provides a clearer view of the MSF leadership field.

Examining the results of the EFA process in this study made the contradictions in the literature about the impact of rater source and dimensions on MSF outcomes more clear. Asking

identical questions across rater groups and then comparing CFA results is not a sound practice when the EFA process proves a large number of the items do not even load for all rater groups. EFA analyses also supported why self vs. other results are inconsistent and hard to compare. The same factors emerge, but the items comprising those factors and the amount of variance explained by those factors differ a great deal.

These item and factor inconsistencies are what MI is designed to detect, but the lack of data exploration and MI testing in the literature adds even more to our understanding of why so many discrepancies exist in MSF research results. If the most popular and widely used leadership development instruments do not adhere to industry best practice methods standards, it is evident why the findings lack consistency. More importantly, it highlights a significant defect in the MSF research because without MI testing the results that are reported cannot be properly interpreted (Horn and McArdle, 1992).

The idea of flawed rater group results applied in organizations is disturbing. A comprehensive look at the MSF leadership tool market makes it easy to understand how and why organizations spend so much time and money on the process. Hundreds of companies compete to sell and administer MSF tests, analyze the data and deliver the results. If the assessment was for developmental purposes, then additional and significant amounts of money are spent developing, administering and overseeing coaching plans. If the assessment is for administrative and/or performance evaluation purposes, the time and money investments take on even greater significance as implementing the perceived results can have a profound impact on organizational outcomes such as profitability, leadership development, leadership succession, pay and promotion. It is a serious concern that MSF leadership tools are not properly validated, the data are not properly explored and the results are not properly supported by required MI testing.

The MSF leadership field has struggled with inconsistencies in dimensions of interest, models of raters and concrete results since its inception. Until there is an agreement in the literature as to instrument use and validation, models of MSF leadership and the appropriateness of the application of the results, the field will continue to suffer from conflicting and unsubstantiated results. More importantly, even though comparing leadership feedback from different sources might provide the most accurate depiction of the leader, organizations need to decide if the significant amounts of time and money devoted to the process yield results that are worth the investment.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin,* 103, 411-423.

Atwater, L., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self- and follower ratings. *Personnel Psychology, 48*(1), 35-59.

Atwater, L., Waldman, D., Ostroff, C., Robie, C., & Johnson, K. M. (2005). Self-other agreement: Comparing its relationship with performance in the U.S. and Europe. *International Journal of Selection and Assessment, 13*(1), 25-40.

Atwater, L. E. & Brett, J. F. (2006). 360-Degree feedback to leaders: Does it relate to changes in employee attitudes? *Group & Organization Management, 31*(5), 578-600.

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement: Does it really matter? *Personnel Psychology, 51*(3), 577-598.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods, 1*(1), 45-87.

Barr, M. A., & Raju, N. S. (2003). IRT-Based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods, 6*(1), 15-44.

Bartol, K. M., Martin, D. C., & Kromkowski, J. A. (2003). Leadership and the glass ceiling: Gender and ethnic group influences on leader behaviors at middle and executive managerial levels. *Journal of Leadership & Organizational Studies, 9*(3), 8-19.

Beatty, R. (2005). Leadership Development. In *The Blackwell Encyclopedia of Management: Human Resource Management* (Vol. 5, p.). Malden, MA: Blackwell Publishing.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin 107*(2): 238-246.

Bentler, P. M. & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin***,** *88*(3): 588-606.

Borman, W. C. (1974). The rating of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance 12*(1): 105-124.

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*(3), 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), 299-316.

Borman, W. C. (2005). Peer Ratings. In *The Blackwell Encyclopedia of Management: Human Resource Management* (Vol. 5, p.). Malden, MA: Blackwell Publishing.

Borman, W. C. (2005). Self-ratings. In *The Blackwell Encyclopedia of Management: Human Resource Management* (Vol. 5, p. 340). Malden, MA: Blackwell Publishing.

Borman, W. C. (2005). Supervisor Ratings. In *The Blackwell Encyclopedia of Management: Human Resource Management* (Vol. 5, p.). Malden, MA: Blackwell Publishing.

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance, 6*(1), 1-21.

Borman, W. C. & Motowidlo, S.J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), Personnel Selection in organizations (pp. 71-98). San Francisco: Jossey-Bass.

Bryman, A. (1987). The Generalizability of Implicit Leadership Theory. *Journal of Social Psychology, 127*(2), 129.

Burnette, J. L. & Williams, L. J. (2005). Structural equation modeling (SEM): An introduction to basic techniques and advanced issues. In R. A. Swanson & E. F. Holton III (Eds.), Research in organizations : Foundations and methods of inquiry (pp. 143-160). San Francisco: Berrett-Koehler.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods & Research, 5,* 3-52.

Cascio, W. F., & Aguinis, H. (2008). Staffing Twenty-first-century Organizations. *Academy of Management Annals, 2*(1), 133-165.

Cheung, G. W. (1999). Multifaceted conceptions of self-other ratings disagreement. *Personnel Psychology, 52*(1), 1-36.

Cheung, G.W. & Rensvold, R.B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new model. *Journal of Management, 25,* 1-27.

Cheung, G. W. & Rensvold**,** R. B. (2002). Evaluating Goodness of fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255.

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erbaum.

Conway, J. M. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology, 84*(1), 3-13.

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*(4), 331-360.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Craig, S. B., & Kaiser, R. B. (2003). Applying item response theory to multisource performance ratings: What are the consequences of violating the independent observations assumption? *Organizational Research Methods, 6*(1), 44-61.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika. 16*, 297-334.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior & Human Performance, 33*(3), 360-396.

Diefendorff, J. M., Silverman, S. B., & Greguras, G. J. (2005). Measurement equivalence and multisource ratings for non-managerial positions: Recommendations for research and practice. *Journal of Business & Psychology. 19*(3), 399-425.

Dierdorff, E. C., & Surface, E. A. (2007). Placing peer ratings in context: Systematic Influences beyond rate performance. *Personnel Psychology,60*(2), 93-126.

Dierdorff, E. C., Surface, E. A., Meade, A., Thompson, L. F., & Martin, D. L. (2006). Group differences and measurement equivalence: Implications for command climate survey research and practice. *Military Psychology, 18*(1), 19-37.

Eden, D., & Leviatan, U. (1975). Implicit Leadership Theory as a Determinant of the Factor Structure Underlying Supervisory Behavior Scales. *Journal of Applied Psychology, 60*(6), 736-741.

Epitropaki, O., & Martin, R. (2004). Implicit leadership theories in applied settings: factor structure, generalizability, and stability over time. *Journal of Applied Psychology, 89*(2), 293-310.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology, 86*(2), 215-227.

Facteau, C. L. & Facteau, J. D. (1998). Reactions of leaders to 360-degree feedback from subordinates and peers. *Leadership Quarterly, 9*(4), 427-449.

Field, A. (2011). *Discovering statistics using SPSS. (5th ed.)*. London: SAGE.

Fielding, K. S., & Hogg, M. A. (1997). Social identity, self-categorization, and leadership: A field study of small interactive groups. *Group Dynamics: Theory, Research, and Practice, 1*(1), 39-51.

Foti, R. J., Fraser, S. L., & Lord, R. G. (1982). Effects of leadership labels and prototypes on perceptions of political leaders. *Journal of Applied Psychology, 67*(3), 326-333.

Fox, S., & Bizman, A. (1988). Differential Dimensions Employed in Rating Subordinates, Peers, and Superiors. *Journal of Psychology*. *122*(4), 373-473.

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology, 83*(6), 960-968.

Greguras, G. J., Robie, C., Schleicher, D. J., & Goff, M. (2003). A field study of the effects of rating purpose on the quality of multisource ratings. *Personnel Psychology, 56*(1), 1-21.

Hains, S. C., Hogg, M. A., & Duck, J. M. (1997). Self-categorization and leadership: Effects of group prototypicality and leader stereotypicality. *Personality and Social Psychology Bulletin, 23*(10), 1087-1099.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*(1), 43-62.

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology. 63*(1), *119-151.*

Hooijberg, R., & Choi, J. (2000). Which leadership roles matter to whom? An examination of rater effects on perceptions of effectiveness. *Leadership Quarterly, 11*(3), 341-365.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3-4), 117-144.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling 6*(1): 1-55.

Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., Williams, L.J. (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior, 18* (1), 667-683.

James, L. R., Demaree, R., G. &Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*, 85-98.

Johnson, E.C., & Meade, A.W. (2010, April). *A Multi-Level Investigation of Overall Job Performance Ratings.* Paper presented at the 25th Annual Meeting of the Society for

Industrial and Organizational Psychology, Atlanta, GA.

Kaiser, R. B., Hogan, R., & Craig, S. B. (2008). Leadership and the fate of organizations. *American Psychologist, 63*(2), 96-110.

Katz, D. & Kahn, R.L. (1978). *The social psychology of organizations (2nd ed.)*. New York: Wiley.

Klein, R. (2005). *Principles and practice of structural equation modeling (2nd ed.).* New York: Guildford Press.

Kline, P. (1999). *The handbook of psychological testing (2nd ed.)*. London: Routledge.

Lakey, C., Goodie, A., Lance, C., Stinchfield, R., Winters, K. (2007). Examining DSM-IV criteria for pathological gambling: Psychometric properties and evidence from cognitive biases. *Journal of Gambling Studies, 23*(4), 479-498.

Lance, C. E. (1994). Test of a Latent Structure of Performance Ratings Derived from Wherry's (1952) Theory of Rating, *Journal of Management, 20*(40), 757-771.

Lance, C. E., & Bennett, J. W. (2000). Replication and Extension of Models of Supervisory Job Performance Ratings. *Human Performance, 13*(2), 2-22.

Landy, F. J., Farr, J. L., Saal, F. E., & Freytag, W. R. (1976). Behaviorally anchored scales for rating the performance of police officers. *Journal of Applied Psychology, 61*(6), 750-758.

Lawler, E. E. (1967). The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology, 51*(5), 369-381.

LeBreton, J.M. (2011, October). Email exchange of ICC's in group research, personal electronic communication. Richmond, VA.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*(1), 80-129.

LeBreton, J.M. & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4) 815-853.

Leslie, J.B., & Fleenor, J.W. (1998). *Feedback to managers: A review and comparison of multi-rater instruments for management development.* Greensboro, NC: Center for Creative Leadership.

London, M., & Smither, J. W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self evaluations, and performance-related outcomes? Theory-based

applications and directions for research. *Personnel Psychology*, *48*(4), 803-839.

Lord, R. G., Brown, D. J., & Freiberg, S. J. (1999). Understanding the dynamics of leadership: The role of follower self-concepts in the leader/follower relationship. *Organizational Behavior and Human Decision Processes*, *78*(3), 167-203.

Lord, R. G., Brown, D. J., Harvey, J. L., & Hall, R. J. (2001). Contextual constraints on prototype generation and their multilevel consequences for leadership perceptions. *Leadership Quarterly*, *12*(3), 311-339.

Lord, R. G., Foti, R. J., & De Vader, C. L. (1984). A Test of Leadership Categorization Theory: Internal Structure, Information Processing, and Leadership Perceptions. *Organizational Behavior & Human Performance*, *34*(3), 343-378.

Mabe, P. A., & West, S. G. (1982). Validity of Self-Evaluation of Ability: A Review and Meta-Analysis. *Journal of Applied Psychology*, *67*(3), 280-296.

Mann, F. C. (1965). *Toward an understanding of the leadership role in formal organizations*. In R. Dubin, G.C. Homans, F.C. Mann, & D.C. Miller (Eds.), Leadership and productivity. San Francisco: Chandler.

Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology*, *83*(5), 693-702.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling-a Multidisciplinary Journal*, *14*(4), 611-635.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361-388.

Moses, J., Hollenbeck, G. P., Sorcher, M. (1993). Other people's expectations. *Human Resource Management*, *32* (2/3), 283-297.

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.

Nilsen, D., & Campbell, D. P. (1993). Self-Observer Rating Discrepancies: Once an Overrater, Always an Overrater? *Human Resource Management*, *32*(2/3), 265-281.

Nunnally, J.C., & Bernstein, I.H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

Offermann, L. R., Kennedy, J. K., & Wirtz, P. W. (1994). Implicit leadership theories: Content,

structure, and generalizability. *The Leadership Quarterly*, *5*(1), 43-58.

Organizational Research Methods. *Most-cited articles: A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research* (downloaded 9/7/2011 and 1/15/2012). http://orm.sagepub.com/reports/most-cited

Ostroff, C., Atwater, L.E., & Feinberg, B.J. (2004). Understanding self-other agreement: A look at rater and ratee characteristics, contexts, and outcomes. *Personnel Psychology*, 57, 333-375.

Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of Job Performance Ratings: An Examination of Ratee Race, Ratee Gender, and Rater Level Effects. *Human Performance, 9*(2), 103-120.

Ritter, B. A., & Lord, R. G. (2007). The impact of previous leaders on the evaluation of new leaders: An alternative to prototype matching. *Journal of Applied Psychology, 92*(6), 1683-1695.

Roberson, L., Galvin, B. M., & Charles, A. C. (2007). When Group Identities Matter: Bias in Performance Appraisal. *Academy of Management Annals. 1*(13), 617-650.

Rosch, E. (1975). The nature of mental codes for color categories. *Journal of Experimental Psychology: Human Perception and Performance, 1*(4), 303-322.

Rosette, A. S., Leonardelli, G. J., & Phillips, K. W. (2008). The White standard: Racial bias in leader categorization. *Journal of Applied Psychology, 93*(4), 758-777.

Roth, P. L. and Switzer, F. S. (1995). "A Monte Carlo Analysis of missing data techniques in a HRM setting." *Journal of Management, 21*(5): 1003-1023.

Roth, P. L., Switzer, F. S., Switzer D.M. (1999). Missing data in multiple item scales: A Monte Carlo Analysis of missing data techniques." *Organizational Research Methods 2*(3), 211-233

Rush, M C, Thomas, J C, & Lord, R G. (1977). Implicit leadership theory: A potential threat to the internal validity of leader behavior questionnaires. *Organizational Behavior and Human Performance, 20*(1), 93-110.

Salam, S., Cox, J. F., & Sims, J., Henry P. (1997). In the eye of the beholder: How leadership relates to 360-degree performance ratings. *Group & Organization Management, 22*(2), 185-209.

Schoorman, F. D., & Mayer, R. C. (2008). The value of common perspectives in self-reported appraisals: You get what you ask for. *Organizational Research Methods, 11*(1), 148-159.

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

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.

Smither, J. W., Brett, J. F., & Atwater, L. (2008). What Do Leaders Recall About Their Multisource Feedback? *Journal of Leadership & Organizational Studies, 14*(3), 202-218.

Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology, 58*(1), 33-66.

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

Spangler, M. (2003). Review of Benchmarks [revised]. In B.S. Plake, J. Impara, & R.A. Spies (Eds.), The fifteenth mental measurements yearbook (pp. 124-126). Lincoln, NE: Buros Institute of Mental Measurements.

Spector, P. (2006). Method Variance in Organizational Research: Truth or Urban Legend? *Organizational Research Methods, 9*(2), 221-232.

Stephenson, W. (1935). Technique of factor analysis. Nature, 136, 297

Stephenson, W. (psychologist). (2011, September 12). In *Wikipedia, The Free Encyclopedia*. Retrieved 20:21, November 9, 2011, from http://en.wikipedia.org/w/index.php?title=William_Stephenson_(psychologist)&oldid=449936580.

Tsui, A. S., & Ohlott, P. (1988). Multiple assessment of managerial effectiveness: Interrater agreement and consensus in effectiveness models. *Personnel Psychology, 41*(4), 779-803.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158.

Vandenberg, R. J. (2010, May). *Structural Equation Modeling*. Center for the Advancement of Research Methods and Analysis (CARMA) Short Course entitled Structural Equation Modeling, R. Vandenberg, Instructor. Richmond, VA.

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*(1), 4-69.

Vandenberg, R. J., & Self, R. M. (1993). Assessing newcomers' changing commitments to the organization during the first 6 months of work. *Journal of Applied Psychology, 78*(4), 557-568.

vanDierendonck, D., Haynes, C., Borrill, C., & Stride, C. (2007). Effects of upward feedback on leadership behaviour toward subordinates. *The Journal of Management Development, 26*(3), 228-238.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (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. (2002). The moderating influence of job performance dimensions on convergence of supervisory and peer ratings of job performance: Unconfounding construct-level convergence and rating difficulty. *Journal of Applied Psychology, 87*(2), 345-354.

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*(1), 108-131.

Waldman, D. A., & Atwater, L. E. (2001). Attitudinal and behavioral outcomes of an upward feedback process. *Group & Organization Management, 26*(2), 189-205.

Williams, L. J. & O'Boyle, E. H. (2008) Measurement models for linking latent variables and indicators: A review of human resource management research using parcels. *Human Resource Management Review, 18*(4), 233-242.

Williams, L. J., Vandenberg, R. J., & Edwards, J. R. (2009). Structural equation modeling in management research: A guide for improved analysis. *Academy of Management Annals, 3*(1), 543-604.

Woehr, D. J., Sheehan, M. K., & Bennett, J. W. (2005). Assessing Measurement Equivalence Across Rating Sources: A Multitrait-Multirater Approach *Journal of Applied Psychology, 90(3)*, 592-600.

Table 1a. Self Group: Original Item Descriptives including Sample Size, Means, Standard Deviations, Missing Data and Extremes

Self	N	Mean	Std. Deviation	Missing	
				Count	Percent
Results					
1-Res	4467	4.01	.55	10	.2
2-Res	4469	4.22	.52	8	.2
3-Res	4461	4.10	.50	16	.4
4-Res	4469	4.30	.52	8	.2
Decisiveness					
1-Dec	4470	3.98	.56	7	.2
2-Dec	4465	4.18	.57	12	.3
3-Dec	4468	4.01	.57	9	.2
Strategic Focus					
1-Str	4463	4.19	.52	14	.3
2-Str	4467	3.91	.53	10	.2
3-Str	4467	4.07	.53	10	.2
4-Str	4367	3.75	.71	110	2.5
5-Str	4432	3.72	.60	45	1.0
6-Str	4450	3.34	.77	27	.6
7-Str	4464	4.02	.55	13	.3
8-Str	4403	3.87	.63	74	1.7
9-Str	4407	3.92	.70	70	1.6
Courageous Authenticity					
1-Cou	4467	3.83	.67	10	.2
2-Cou	4464	3.92	.65	13	.3
3-Cou	4466	3.87	.66	11	.2
Integrity					
1-Int	4467	4.31	.47	10	.2
2-Int	4466	4.38	.50	11	.2
3-Int	4456	4.18	.53	21	.5
Leadership Effectiveness					
1-LEEff	4464	3.61	.65	13	.3
2-LEEff	4416	3.50	.65	61	1.4
3-LEEff	4442	3.34	.66	35	.8
4-LEEff	4441	4.01	.56	36	.8
5-LEEff	4453	3.93	.53	24	.5
Caring Connection					
1-Car	4466	3.58	.78	11	.2
2-Car	4467	4.13	.61	10	.2
3-Car	4469	3.94	.69	8	.2
Collaborator					
1-Col	4461	3.98	.57	16	.4
2-Col	4464	4.06	.52	13	.3
3-Col	4469	4.00	.53	8	.2
Fosters Team Play					
1-Fos	4460	4.14	.52	17	.4
2-Fos	4449	3.97	.62	28	.6
3-Fos	4464	4.03	.59	13	.3

Interpersonal Intelligence					
1-Int	4468	4.22	.52	9	.2
2-Int	4463	3.76	.63	14	.3
3-Int	4468	3.88	.60	9	.2
4-Int	4462	3.86	.67	15	.3
5-Int	4468	3.77	.64	9	.2
Mentoring and Developing					
1-**Men**	**4143**	**3.72**	**.74**	**334**	**7.5**
2-Men	4458	3.92	.67	19	.4
3-Men	4443	3.91	.59	34	.8
4-Men	4468	4.12	.55	9	.2
Systems Thinker					
1-Sys	4450	3.83	.62	27	.6
2-Sys	4377	3.59	.68	100	2.2
3-Sys	4353	3.75	.66	124	2.8

Bold indicates the item that was deleted after initial data screening.

Table 1b. Supervisor Group: Original Item Descriptives including Sample Size, Means, Standard Deviations, Missing Data and Extremes

Supervisor	N	Mean	Std. Deviation	Missing	
				Count	Percent
Results					
1-Res	4963	4.08	.66	16	.3
2-Res	4968	4.30	.63	11	.2
3-Res	4954	4.17	.60	25	.5
4-Res	4967	4.22	.60	12	.2
Decisiveness					
1-Dec	4969	4.00	.65	10	.2
2-Dec	4940	4.03	.72	39	.8
3-Dec	4957	4.06	.67	22	.4
Strategic Focus					
1-Str	4954	4.02	.69	25	.5
2-Str	4945	3.87	.66	34	.7
3-Str	4956	4.02	.65	23	.5
4-Str	4739	3.96	.72	240	4.8
5-Str	4870	3.82	.72	109	2.2
6-Str	4840	3.51	.83	139	2.8
7-Str	4951	3.98	.67	28	.6
8-Str	4852	3.88	.74	127	2.6
9-Str	4821	4.06	.72	158	3.2
Courageous Authenticity					
1-Cou	4951	3.81	.78	28	.6
2-Cou	4951	3.99	.72	28	.6
3-Cou	4947	3.93	.74	32	.6
Integrity					
1-Int	4950	4.41	.55	29	.6
2-Int	4925	4.38	.56	54	1.1
3-Int	4895	4.26	.58	84	1.7
Leadership Effectiveness					
1-LEEff	4954	4.10	.71	25	.5
2-LEEff	4917	3.87	.78	62	1.2
3-LEEff	4936	3.75	.76	43	.9
4-LEEff	4929	4.15	.69	50	1.0
5-LEEff	4931	4.08	.70	48	1.0
Caring Connection					
1-Car	4856	3.61	.82	123	2.5
2-Car	4952	4.12	.69	27	.5
3-Car	4940	3.96	.79	39	.8
Collaborator					
1-Col	4937	3.95	.66	42	.8
2-Col	4955	4.00	.67	24	.5
3-Col	4958	3.93	.68	21	.4
Fosters Team Play					
1-Fos	4959	4.09	.69	20	.4
2-Fos	4938	3.92	.77	41	.8
3-Fos	4934	3.95	.72	45	.9

Interpersonal Intelligence					
1-Int	4909	4.03	.72	70	1.4
2-Int	4872	3.80	.76	107	2.1
3-Int	4948	4.00	.73	31	.6
4-Int	4926	3.76	.81	53	1.1
5-Int	4930	3.77	.79	49	1.0
Mentoring and Developing					
1-Men	**4347**	**3.83**	**.77**	**632**	**12.7**
2-Men	4882	3.85	.79	97	1.9
3-Men	4730	3.89	.69	249	5.0
4-Men	4951	4.07	.66	28	.6
Systems Thinker					
1-Sys	4863	3.81	.74	116	2.3
2-Sys	4782	3.66	.75	197	4.0
3-Sys	4774	3.81	.74	205	4.1

Bold indicates the item that was deleted based on initial data screening.

Table 1c. Direct Report Group: Original Item Descriptives including Sample Size, Means, Standard Deviations, Missing Data and Extremes

Direct Report	N	Mean	Std. Deviation	Missing	
				Count	Percent
Results					
1-AR	16632	4.24	.69	94	.6
2-AR	16669	4.37	.66	57	.3
3-AR	16544	4.26	.67	182	1.1
4-AR	16657	4.35	.67	69	.4
Decisiveness					
1-Dec	16678	4.15	.77	48	.3
2-Dec	16586	4.25	.77	140	.8
3-Dec	16661	4.06	.80	65	.4
Strategic Focus					
1-Str	16572	4.21	.74	154	.9
2-Str	16493	3.97	.77	233	1.4
3-Str	16667	4.21	.70	59	.4
4-Str	15908	4.22	.71	818	4.9
5-Str	16560	3.96	.82	166	1.0
6-Str	16214	3.83	.87	512	3.1
7-Str	16589	4.10	.74	137	.8
8-Str	16540	4.16	.79	186	1.1
9-Str	16059	4.26	.71	667	4.0
Courageous Authenticity					
1-Cou	16476	3.85	.88	250	1.5
2-Cou	16555	4.07	.80	171	1.0
3-Cou	16432	4.14	.77	294	1.8
Integrity					
1-Int	16501	4.39	.66	225	1.3
2-Int	16392	4.36	.64	334	2.0
3-Int	16252	4.28	.68	474	2.8
Leadership Effectiveness					
1-LEEff	16682	4.195	.83	44	.3
2-LEEff	16610	3.98	.94	116	.7
3-LEEff	16650	3.93	.91	76	.5
4-LEEff	16638	4.26	.79	88	.5
5-LEEff	16678	4.20	.83	48	.3
Caring Connection					
1-Car	16172	3.50	.97	554	3.3
2-Car	16632	4.09	.83	94	.6
3-Car	16594	3.96	.94	132	.8
Collaborator					
1-Col	16538	3.97	.80	188	1.1
2-Col	16624	4.02	.78	102	.6
3-Col	16654	4.02	.78	72	.4
Fosters Team Play					
1-Fos	16686	4.16	.84	40	.2
2-Fos	16629	3.95	.93	97	.6
3-Fos	16598	4.01	.83	128	.8

Interpersonal Intelligence					
1-Int	16072	4.04	.85	654	3.9
2-Int	16004	3.85	.88	722	4.3
3-Int	16370	3.94	.89	356	2.1
4-Int	16533	3.78	1.01	193	1.2
5-Int	16541	3.94	.89	185	1.1
Mentoring and Developing					
1-Men	**15708**	**3.88**	**.93**	**1018**	**6.1**
2-Men	16609	3.88	.96	117	.7
3-Men	16453	3.94	.89	273	1.6
4-Men	16663	4.13	.83	63	.4
Systems Thinker					
1-Sys	16288	3.74	.91	438	2.6
2-Sys	15817	3.62	.86	909	5.4
3-Sys	15987	3.91	.79	739	4.4

Bold indicates the item that was deleted based on initial data screening.

Table 1d. Peer Group: Original Item Descriptives including Sample Size, Means, Standard Deviations, Missing Data and Extremes

Peer	N	Mean	Std. Deviation	Missing	
				Count	Percent
Results					
1-Res	19869	4.09	.67	117	.6
2-Res	19927	4.22	.65	59	.3
3-Res	19732	4.14	.64	254	1.3
4-Res	19878	4.20	.64	108	.5
Decisiveness					
1-Dec	19871	4.03	.70	115	.6
2-Dec	19706	4.05	.73	280	1.4
3-Dec	19788	4.02	.70	198	1.0
Strategic Focus					
1-Str	19787	4.04	.73	199	1.0
2-Str	19666	3.88	.70	320	1.6
3-Str	19874	4.04	.67	112	.6
4-Str	18806	4.02	.72	1180	5.9
5-Str	19530	3.85	.75	456	2.3
6-Str	19218	3.60	.84	768	3.8
7-Str	19754	4.01	.69	232	1.2
8-Str	19464	3.94	.76	522	2.6
9-Str	19081	4.13	.70	905	4.5
Courageous Authenticity					
1-Cou	19751	3.79	.82	235	1.2
2-Cou	19842	3.98	.76	144	.7
3-Cou	19752	3.97	.77	234	1.2
Integrity					
1-Int	19766	4.35	.60	220	1.1
2-Int	19568	4.29	.61	418	2.1
3-Int	19434	4.20	.63	552	2.8
Leadership Effectiveness					
1-LEEff	19800	4.08	.78	186	.9
2-LEEff	19751	3.81	.86	235	1.2
3-LEEff	19826	3.73	.83	160	.8
4-LEEff	19683	4.10	.74	303	1.5
5-LEEff	19808	4.05	.76	178	.9
Caring Connection					
1-Car	19341	3.57	.89	645	3.2
2-Car	19856	4.06	.79	130	.7
3-Car	19789	3.95	.86	197	1.0
Collaborator					
1-Col	19700	3.91	.72	286	1.4
2-Col	19865	3.97	.73	121	.6
3-Col	19897	3.93	.73	89	.4
Fosters Team Play					
1-Fos	19839	4.06	.76	147	.7
2-Fos	19704	3.87	.84	282	1.4
3-Fos	19716	3.95	.77	270	1.4

Interpersonal Intelligence					
1-Int	19326	3.98	.78	660	3.3
2-Int	19262	3.81	.79	724	3.6
3-Int	19596	3.89	.79	390	2.0
4-Int	19505	3.75	.87	481	2.4
5-Int	19661	3.81	.82	325	1.6
Mentoring and Developing					
1-Men	**15247**	**3.84**	**.80**	**4739**	**23.7**
2-Men	19368	3.81	.87	618	3.1
3-Men	18361	3.86	.78	1625	8.1
4-Men	19700	4.04	.75	286	1.4
Systems Thinker					
1-Sys	19100	3.76	.80	886	4.4
2-Sys	18622	3.63	.78	1364	6.8
3-Sys	18854	3.81	.75	1132	5.7

Bold indicates the item that was deleted based on initial data screening.

Table 1e. Supervisor's Boss Group: Original Item Descriptives including Sample Size, Means, Standard Deviations, Missing Data and Extremes

Supervisor's Boss	N	Mean	Std. Deviation	Missing	
				Count	Percent
Results					
1-Res	1949	4.02	.68	20	1.0
2-Res	1966	4.26	.64	3	.2
3-Res	1948	4.14	.62	21	1.1
4-Res	1960	4.19	.63	9	.5
Decisiveness					
1-Dec	1943	3.92	.70	26	1.3
2-Dec	1927	3.98	.73	42	2.1
3-Dec	1933	3.98	.70	36	1.8
Strategic Focus					
1-Str	1950	3.95	.74	19	1.0
2-Str	1934	3.82	.69	35	1.8
3-Str	1951	3.97	.67	18	.9
4-Str	1851	3.92	.72	118	6.0
5-Str	1875	3.77	.76	94	4.8
6-Str	1864	3.44	.83	105	5.3
7-Str	1927	3.93	.69	42	2.1
8-Str	1879	3.84	.78	90	4.6
9-Str	1866	4.04	.73	103	5.2
Courageous Authenticity					
1-Cou	1932	3.73	.78	37	1.9
2-Cou	1954	3.92	.74	15	.8
3-Cou	1948	3.85	.77	21	1.1
Integrity					
1-Int	1956	4.38	.59	13	.7
2-Int	1922	4.34	.57	47	2.4
3-Int	1923	4.20	.63	46	2.3
Leadership Effectiveness					
1-LEEff	1954	4.02	.75	15	.8
2-LEEff	1924	3.78	.81	45	2.3
3-LEEff	1942	3.66	.79	27	1.4
4-LEEff	1948	4.09	.74	21	1.1
5-LEEff	1945	3.99	.73	24	1.2
Caring Connection					
1-Car	1908	3.58	.82	61	3.1
2-Car	1953	4.11	.72	16	.8
3-Car	1940	3.94	.81	29	1.5
Collaborator					
1-Col	1931	3.90	.69	38	1.9
2-Col	1946	3.95	.71	23	1.2
3-Col	1949	3.88	.72	20	1.0
Fosters Team Play					
1-Fos	1946	4.03	.74	23	1.2
2-Fos	1925	3.88	.81	44	2.2
3-Fos	1929	3.91	.73	40	2.0

Interpersonal Intelligence					
1-Int	1907	3.97	.72	62	3.1
2-Int	1879	3.76	.78	90	4.6
3-Int	1938	3.91	.74	31	1.6
4-Int	1915	3.71	.83	54	2.7
5-Int	1927	3.69	.82	42	2.1
Mentoring and Developing					
1-Men	**1553**	**3.83**	**.76**	**416**	**21.1**
2-Men	1891	3.78	.82	78	4.0
3-Men	1765	3.80	.72	204	10.4
4-Men	1936	4.02	.70	33	1.7
Systems Thinker					
1-Sys	1889	3.75	.75	80	4.1
2-Sys	1829	3.62	.79	140	7.1
3-Sys	1838	3.79	.74	131	6.7

Bold indicates the item that was deleted based on initial data screening.

Table 2 - Exploratory Factor Analysis Results for 42 Items-Principal Component Factoring with Oblimin Rotation

Dimension	1					2					3				
Rater Group*	**Self**	**Boss**	**D.R.**	**Peer**	**Sup. Boss**	**Self**	**Boss**	**D.R.**	**Peer**	**Sup. Boss**	**Self**	**Boss**	**D.R.**	**Peer**	**Sup. Boss**
Achieves Results															
1-Res**															
2-Res	.541											.549			
3-Res															
4-Res															
Decisiveness															
1-Dec	.641														
2-Dec	.742														.645
3-Dec	.724														.556
Strategic Focus															
1-Str	.628	.593	.601												.589
2-Str	.567	.620	.627									.426			
3-Str															
4-Str						**.853**	**.895**	**.871**	**.761**		**.677**				
5-Str						**.734**	**.625**	**.730**	**.663**		**.606**				
6-Str						**.782**	**.675**	**.723**	**.776**		**.770**				
7-Str						**.695**	**.640**	**.749**	**.664**		**.640**				
8-Str						**.665**	**.596**	**.624**	**.648**		**.668**				
9-Str						**.819**	**.864**	**.830**	**.728**		**.638**				
Courag Authent															
1-Cou	**.638**											**.660**	**.774**	**.858**	**.705**
2-Cou	**.698**											**.861**	**.700**	**.832**	**.824**
3-Cou	**.519**											**.782**	**.621**	**.826**	**.793**
Integrity															
1-Int															
2-Int	.561														
3-Int															
Caring Connect.															
1-Car						**.695**	**.814**	**.875**	**.873**	**.795**					
2-Car						**.704**	**.846**	**.925**	**.948**	**.855**					
3-Car						**.763**	**.921**	**.954**	**.950**	**.895**					
Collaborator															
1-Col						**.528**	**.722**	**.641**	**.704**	**.702**					
2-Col						**.674**	**.810**	**.742**	**.794**	**.804**					
3-Col						**.679**	**.774**	**.718**	**.772**	**.793**					
Foster Team Play															
1-Fos						**.672**	**.768**	**.723**	**.770**	**.790**					
2-Fos						**.603**	**.742**	**.631**	**.710**	**.719**					
3-Fos						**.581**	**.706**	**.667**	**.714**	**.743**					
Int. Intelligence															
1-Int										.741					
2-Int											.551				
3-Int											.657				
4-Int	**.588**											**.584**	**.687**	**.547**	**.444**
5-Int											.630				
Mentor/Develop															
1-Men						**.659**	**.728**	**.704**	**.753**	**.736**					
2-Men											.486				
3-Men						**.630**	**.623**	**.614**	**.686**	**.666**					
Systems Thinker															
1-Sys															
2-Sys			**.669**	**.546**	**.710**	**.739**					**.635**				
3-Sys			**.569**	**.581**	**.672**	**.698**					**.574**				

*By Rater Group:	Self	Supervisor	Direct Report	Peer	Supervisor's Boss
Eigenvalue	18.6	25	26	26.4	26
Total Variance Explained	**44.2%**	**59.5%**	**61.9%**	**62.9%**	**62.1%**

Items in black bold loaded for all rater groups = 23 items
Dimension 1: 8 items
Dimension 2: 11 items
Dimension 3: 4 items

Table 3 - Scale Reliabilities for EFA Dimensions and Leadership Effectiveness

Rater Group	Analytical (8 items)	Interpersonal (11 items)	Courageous (4 items)	Leadership Effectiveness (5 items)
Leader	**.85**	**.88**	**.79**	**.87**
Supervisor	**.92**	**.94**	**.83**	**.96**
Direct Report	**.93**	**.95**	**.80**	**.97**
Peer	**.93**	**.95**	**.83**	**.96**
Supervisor's Boss	**.93**	**.95**	**.84**	**.96**

Table 4 - Correlations by Rater Groups, Items and Dimensions

Analytical Dimension

Leader	Ana1	Ana2	Ana3	Ana4	Ana5	Ana6	Ana7	Ana8
Ana1	1	.393**	.408**	.357**	.407**	.594**	.314**	.326**
Ana2	.393**	1	.435**	.498**	.565**	.362**	.415**	.449**
Ana3	.408**	.435**	1	.474**	.520**	.348**	.419**	.404**
Ana4	.357**	.498**	.474**	1	.530**	.356**	.452**	.443**
Ana5	.407**	.565**	.520**	.530**	1	.415**	.437**	.538**
Ana6	.594**	.362**	.348**	.356**	.415**	1	.285**	.325**
Ana7	.314**	.415**	.419**	.452**	.437**	.285**	1	.533**
Ana8	.326**	.449**	.404**	.443**	.538**	.325**	.533**	1
Supervisor								
Ana1	1	.558**	.585**	.519**	.565**	.669**	.462**	.484**
Ana2	.558**	1	.691**	.681**	.755**	.550**	.586**	.641**
Ana3	.585**	.691**	1	.647**	.713**	.548**	.559**	.565**
Ana4	.519**	.681**	.647**	1	.653**	.502**	.582**	.591**
Ana5	.565**	.755**	.713**	.653**	1	.559**	.570**	.635**
Ana6	.669**	.550**	.548**	.502**	.559**	1	.448**	.481**
Ana7	.462**	.586**	.559**	.582**	.570**	.448**	1	.655**
Ana8	.484**	.641**	.565**	.591**	.635**	.481**	.655**	1
Direct Report								
Ana1	1	.586**	.618**	.573**	.611**	.663**	.486**	.551**
Ana2	.586**	1	.692**	.705**	.758**	.535**	.574**	.663**
Ana3	.618**	.692**	1	.656**	.716**	.566**	.583**	.632**
Ana4	.573**	.705**	.656**	1	.697**	.524**	.567**	.620**
Ana5	.611**	.758**	.716**	.697**	1	.573**	.572**	.685**
Ana6	.663**	.535**	.566**	.524**	.573**	1	.455**	.525**
Ana7	.486**	.574**	.583**	.567**	.572**	.455**	1	.634**
Ana8	.551**	.663**	.632**	.620**	.685**	.525**	.634**	1
Peer								
Ana1	1	.587**	.603**	.563**	.599**	.683**	.506**	.535**
Ana2	.587**	1	.700**	.705**	.755**	.551**	.604**	.664**
Ana3	.603**	.700**	1	.661**	.716**	.566**	.600**	.626**
Ana4	.563**	.705**	.661**	1	.688**	.538**	.586**	.618**
Ana5	.599**	.755**	.716**	.688**	1	.576**	.592**	.675**
Ana6	.683**	.551**	.566**	.538**	.576**	1	.478**	.516**
Ana7	.506**	.604**	.600**	.586**	.592**	.478**	1	.666**
Ana8	.535**	.664**	.626**	.618**	.675**	.516**	.666**	1
5 Ana1	1	.577**	.610**	.560**	.587**	.687**	.482**	.497**
Ana2	.577**	1	.709**	.695**	.747**	.543**	.607**	.665**
Ana3	.610**	.709**	1	.654**	.725**	.583**	.593**	.597**
Ana4	.560**	.695**	.654**	1	.652**	.550**	.600**	.634**
Ana5	.587**	.747**	.725**	.652**	1	.565**	.604**	.667**
Ana6	.687**	.543**	.583**	.550**	.565**	1	.479**	.522**
Ana7	.482**	.607**	.593**	.600**	.604**	.479**	1	.685**
Ana8	.497**	.665**	.597**	.634**	.667**	.522**	.685**	1

Interpersonal Dimension

Leader	Int1	Int2	Int3	Int4	Int5	Int6	Int7	Int8	Int9	Int10	Int11
Int1	1	.517**	.643**	.262**	.341**	.277**	.354**	.331**	.256**	.430**	.379**
Int2	.517**	1	.542**	.306**	.379**	.317**	.378**	.328**	.275**	.372**	.341**
Int3	.643**	.542**	1	.255**	.339**	.373**	.439**	.362**	.309**	.430**	.408**
Int4	.262**	.306**	.255**	1	.480**	.450**	.377**	.325**	.321**	.316**	.333**
Int5	.341**	.379**	.339**	.480**	1	.571**	.474**	.488**	.392**	.416**	.420**
Int6	.277**	.317**	.373**	.450**	.571**	1	.501**	.463**	.418**	.406**	.439**
Int7	.354**	.378**	.439**	.377**	.474**	.501**	1	.590**	.415**	.557**	.554**
Int8	.331**	.328**	.362**	.325**	.488**	.463**	.590**	1	.435**	.586**	.525**
Int9	.256**	.275**	.309**	.321**	.392**	.418**	.415**	.435**	1	.376**	.425**
Int10	.430**	.372**	.430**	.316**	.416**	.406**	.557**	.586**	.376**	1	
Int11	.379**	.341**	.408**	.333**	.420**	.439**	.554**	.525**	.425**	.633**	1
Supervisor											
Int1	1	.624**	.727**	.492**	.560**	.524**	.562**	.580**	.472**	.591**	.501**
Int2	.624**	1	.650**	.485**	.557**	.492**	.543**	.523**	.432**	.517**	.463**
Int3	.727**	.650**	1	.523**	.585**	.588**	.617**	.594**	.526**	.575**	.541**
Int4	.492**	.485**	.523**	1	.686**	.668**	.623**	.610**	.557**	.557**	.542**
Int5	.560**	.557**	.585**	.686**	1	.711**	.662**	.690**	.574**	.621**	.583**
Int6	.524**	.492**	.588**	.668**	.711**	1	.660**	.654**	.586**	.589**	.583**
Int7	.562**	.543**	.617**	.623**	.662**	.660**	1	.762**	.605**	.704**	.682**
Int8	.580**	.523**	.594**	.610**	.690**	.654**	.762**	1	.617**	.749**	.667**
Int9	.472**	.432**	.526**	.557**	.574**	.586**	.605**	.617**	1	.574**	.554**
Int10	.591**	.517**	.575**	.557**	.621**	.589**	.704**	.749**	.574**	1	.711**
Int11	.501**	.463**	.541**	.542**	.583**	.583**	.682**	.667**	.554**	.711**1	1
Direct Report											
Int1	1	.677**	.740**	.541**	.593**	.554**	.608**	.615**	.494**	.655**	.581
Int2	.677**	1	.707**	.555**	.626**	.574**	.627**	.581**	.491**	.611**	.561
Int3	.740**	.707**	1	.566**	.620**	.625**	.677**	.628**	.537**	.663**	.633**
Int4	.541**	.555**	.566**	1	.701**	.683**	.684**	.652**	.550**	.627**	.606**
Int5	.593**	.626**	.620**	.701**	1	.733**	.715**	.710**	.588**	.687**	.651**
Int6	.554**	.574**	.625**	.683**	.733**	1	.727**	.689**	.617**	.653**	.661**
Int7	.608**	.627**	.677**	.684**	.715**	.727**	1	.778**	.626**	.751**	.738**
Int8	.615**	.581**	.628**	.652**	.710**	.689**	.778**	1	.621**	.767**	.725**
Int9	.494**	.491**	.537**	.550**	.588**	.617**	.626**	.621**	1	.600**	.600**
Int10	.655**	.611**	.663**	.627**	.687**	.653**	.751**	.767**	.600**	1	.771
Int11	.581**	.561**	.633**	.606**	.651**	.661**	.738**	.725**	.600**	.771**	1
Peer											
Int1	1	.693**	.742**	.540**	.607**	.562**	.621**	.628**	.512**	.649**	.595**
Int2	.693**	1	.710**	.554**	.626**	.573**	.625**	.596**	.518**	.609**	.578**
Int3	.742**	.710**	1	.574**	.627**	.632**	.675**	.637**	.567**	.643**	.632**
Int4	.540**	.554**	.574**	1	.710**	.686**	.676**	.647**	.578**	.605**	.607**
Int5	.607**	.626**	.627**	.710**	1	.736**	.708**	.719**	.615**	.674**	.648**
Int6	.562**	.573**	.632**	.686**	.736**	1	.709**	.682**	.624**	.634**	.651**
Int7	.621**	.625**	.675**	.676**	.708**	.709**	1	.773**	.640**	.737**	.729**
Int8	.628**	.596**	.637**	.647**	.719**	.682**	.773**	1	.646**	.767**	.718**
Int9	.512**	.518**	.567**	.578**	.615**	.624**	.640**	.646**	1	.608**	.617**
Int10	.649**	.609**	.643**	.605**	.674**	.634**	.737**	.767**	.608**	1	.743**
Int11	.595**	.578**	.632**	.607**	.648**	.651**	.729**	.718**	.617**	.743**	1
Supervisor's Boss											
Int1	1	.628**	.714**	.502**	.581**	.547**	.594**	.596**	.504**	.607**	.558**
Int2	.628**	1	.665**	.503**	.580**	.530**	.573**	.541**	.494**	.540**	.516**
Int3	.714**	.665**	1	.545**	.625**	.636**	.653**	.618**	.564**	.615**	.615**
Int4	.502**	.503**	.545**	1	.698**	.661**	.655**	.637**	.571**	.585**	.591**
Int5	.581**	.580**	.625**	.698**	1	.756**	.707**	.721**	.625**	.669**	.651**
Int6	.547**	.530**	.636**	.661**	.756**	1	.707**	.696**	.640**	.653**	.667**
Int7	.594**	.573**	.653**	.655**	.707**	.707**	1	.783**	.672**	.743**	.755**
Int8	.596**	.541**	.618**	.637**	.721**	.696**	.783**	1	.663**	.763**	.717**
Int9	.504**	.494**	.564**	.571**	.625**	.640**	.672**	.663**	1	.592**	.621**
Int10	.607**	.540**	.615**	.585**	.669**	.653**	.743**	.763**	.592**	1	.742**
Int11	.558**	.516**	.615**	.591**	.651**	.667**	.755**	.717**	.621**	.742**	1

Courageous Dimension

Leader	Cou1	Cou2	Cou3	Cou4
Cou1	1	.592**	.567**	.460**
Cou2	.592**	1	.532**	.429**
Cou3	.567**	.532**	1	.343**
Cou4	.460**	.429**	.343**	1
Supervisor				
Cou1	1	.656**	.614**	.504**
Cou2	.656**	1	.655**	.483**
Cou3	.614**	.655**	1	.427**
Cou4	.504**	.483**	.427**	1
Direct Report				
Cou1	1	.612**	.572**	.469**
Cou2	.612**	1	.578**	.446**
Cou3	.572**	.578**	1	.420**
Cou4	.469**	.446**	.420**	1
Peer				
Cou1	1	.660**	.639**	.479**
Cou2	.660**	1	.641**	.465**
Cou3	.639**	.641**	1	.432**
Cou4	.479**	.465**	.432**	1
Supervisor's Boss				
Cou1	1	.655**	.642**	.504**
Cou2	.655**	1	.666**	.469**
Cou3	.642**	.666**	1	.457**
Cou4	.504**	.469**	.457**	1

Leadership Effectiveness

Leader	LEEff1	LEEff2	LEEff3	LEEff4	LEEff5
LEEff1	1	.563**	.586**	.439**	.633**
LEEff2	.563**	1	.698**	.536**	.641**
LEEff3	.586**	.698**	1	.508**	.641**
LEEff4	.439**	.536**	.508**	1	.595**
LEEff5	.633**	.641**	.641**	.595**	1
Supervisor					
LEEff1	1	.820**	.809**	.787**	.867**
LEEff2	.820**	1	.865**	.791**	.847**
LEEff3	.809**	.865**	1	.769**	.831**
LEEff4	.787**	.791**	.769**	1	.819**
LEEff5	.867**	.847**	.831**	.819**	1
Direct Report					
LEEff1	1	.848**	.852**	.812**	.887**
LEEff2	.848**	1	.897**	.828**	.875**
LEEff3	.852**	.897**	1	.822**	.880**
LEEff4	.812**	.828**	.822**	1	.854**
LEEff5	.887**	.875**	.880**	.854**	1
Peer					
LEEff1	1	.828**	.826**	.801**	.867**
LEEff2	.828**	1	.877**	.800**	.855**
LEEff3	.826**	.877**	1	.786**	.853**
LEEff4	.801**	.800**	.786**	1	.832**
LEEff5	.867**	.855**	.853**	.832**	1
Supervisor's Boss					
LEEff1	1	.840**	.828**	.818**	.878**
LEEff2	.840**	1	.866**	.809**	.852**
LEEff3	.828**	.866**	1	.778**	.841**
LEEff4	.818**	.809**	.778**	1	.831**
LEEff5	.878**	.852**	.841**	.831**	1

**Correlation is significant at the 0.01 level (1-tailed).

Table 5 – Group and Dimension Descriptives after Deleting Items Including Mean, Standard Deviation, Minimum, Maximum, and Variance

		ANALYTICAL	INTERPERSONAL	COURAGEOUS	LEADERSHIP EFFECTIVENESS
Group					
Leader					
	Mean	3.76	3.99	3.87	3.68
	N	4177	4427	4455	4385
	Std. Deviation	.46	.41	.52	.50
Supervisor					
	Mean	3.85	3.96	3.88	3.99
	N	4412	4699	4888	4871
	Std. Deviation	.58	.57	.62	.67
Direct Report					
	Mean	4.02	3.98	3.96	4.12
	N	14476	15802	16087	16530
	Std. Deviation	.64	.71	.69	.81
Peer					
	Mean	3.88	3.93	3.87	3.96
	N	16827	18288	19160	19393
	Std. Deviation	.61	.65	.65	.74
Supervisor's Boss					
	Mean	3.82	3.92	3.80	3.92
	N	1618	1784	1877	1902
	Std. Deviation	.61	.61	.64	.71

Table 6 – Dimension Correlations By Rater Group
Dimension/Rater Group

SELF	AnaSelf	InterSelf	CouSelf	LESelf
AnaSelf	1	.41	.54	.60
InterSelf	-	1	.39	.60
CouSelf	-	-	1	.49
LESelf	-	-	-	1

DR	AnaDR	InterDR	Cour DR	LEDR
AnaDR	1	.75	.72	.84
InterDR	.75	1	.64	.87
CouDR	-	-	1	.71
LEDR	-	-	-	1

Super	AnaSup	InterSup	CouSup	LESup
AnaSup	1	.63	.67	.77
InterSup	-	1	.56	.82
CouSup	-	-	1	.69
LESup	-	-	-	1

Peer	AnaPeer	InterPeer	CouPeer	LEPeer
AnaPeer	1	.69	.69	.82
InterPeer	-	1	.57	.85
CouPeer	-	-	1	.68
LEPeer	-	-	-	1

SupBoss	AnaSupBoss	InterSupBoss	CouSupBoss	LESupBoss
AnaSupBoss	1	.65	.69	.79
InterSupBoss	-	1	.58	.84
CouSupBoss	-	-	1	.70
LESupBoss	-	-	-	1

All correlations are significant at the 0.01 level

Table 7 - Games-Howell Post Hoc Tests Between All Rater Groups for Each Dimension

Analytical Dimension

Group(I)	Group(J)	Mean Difference (I-J)	Std. Error	Sig.	99% Confidence Interval Lower Bound	99% Confidence Interval Upper Bound
Self	**Boss**	**-.09481***	**.01130**	**.000**	**-.1316**	**-.0580**
	Dir.Rep	**-.26220***	**.00890**	**.000**	**-.2912**	**-.2332**
	Peer	**-.12754***	**.00856**	**.000**	**-.1554**	**-.0997**
	Sup.Boss	**-.06581***	**.01669**	**.001**	**-.1202**	**-.0114**
Boss	**Self**	**.09481***	**.01130**	**.000**	**.0580**	**.1316**
	Dir.Rep	**-.16740***	**.01021**	**.000**	**-.2006**	**-.1342**
	Peer	**-.03273***	**.00991**	**.009**	**-.0650**	**-.0005**
	Sup.Boss	.02900	.01742	.456	-.0277	.0857
Dir.Rep	**Self**	**.26220***	**.00890**	**.000**	**.2332**	**.2912**
	Boss	**.16740***	**.01021**	**.000**	**.1342**	**.2006**
	Peer	**.13467***	**.00705**	**.000**	**.1117**	**.1576**
	Sup.Boss	**.19640***	**.01596**	**.000**	**.1444**	**.2484**
Peer	**Self**	**.12754***	**.00856**	**.000**	**.0997**	**.1554**
	Boss	**.03273***	**.00991**	**.009**	**.0005**	**.0650**
	Peer	**-.13467***	**.00705**	**.000**	**-.1576**	**-.1117**
	Sup.Boss	**.06173***	**.01577**	**.001**	**.0103**	**.1131**
Sup.Boss	**Self**	**.06581***	**.01669**	**.001**	**.0114**	**.1202**
	Boss	-.02900	.01742	.456	-.0857	.0277
	Dir.Rep	**-.19640***	**.01596**	**.000**	**-.2484**	**-.1444**
	Peer	**-.06173***	**.01577**	**.001**	**-.1131**	**-.0103**

***Bold indicates significant mean difference at the 0.01 level.**

Interpersonal Dimension

Group(I)	Group(J)	Mean Difference (I-J)	Std. Error	Sig.	99% Confidence Interval Lower Bound	Upper Bound
Self	Boss	.03018	.01036	.029	-.0035	.0639
	Dir.Rep	.01296	.00836	.529	-.0142	.0402
	Peer	**.05744***	**.00782**	**.000**	**.0320**	**.0829**
	Sup.Boss	**.06648***	**.01574**	**.000**	**.0152**	**.1178**
Boss	Self	-.03018	.01036	.029	-.0639	.0035
	Dir.Rep	-.01721	.01005	.426	-.0499	.0155
	Peer	.02726	.00961	.037	-.0040	.0585
	Sup.Boss	.03631	.01670	.190	-.0181	.0907
Dir.Rep	Self	-.01296	.00836	.529	-.0402	.0142
	Boss	.01721	.01005	.426	-.0155	.0499
	Peer	**.04447***	**.00741**	**.000**	**.0204**	**.0686**
	Sup.Boss	**.05352***	**.01554**	**.005**	**.0029**	**.1042**
Peer	**Self**	**-.05744***	**.00782**	**.000**	**-.0829**	**-.0320**
	Boss	-.02726	.00961	.037	-.0585	.0040
	Dir.Rep	**-.04447***	**.00741**	**.000**	**-.0686**	**-.0204**
	Sup.Boss	.00905	.01526	.976	-.0407	.0588
Sup.Boss	**Self**	**-.06648***	**.01574**	**.000**	**-.1178**	**-.0152**
	Boss	-.03631	.01670	.190	-.0907	.0181
	Dir.Rep	**-.05352***	**.01554**	**.005**	**-.1042**	**-.0029**
	Peer	-.00905	.01526	.976	-.0588	.0407

***Bold indicates significant mean difference at the 0.01 level.**

Courage Dimension

Group(I)	Group(J)	Mean Difference (I-J)	Std. Error	Sig.	99% Confidence Interval Lower Bound	Upper Bound
Self	Boss	-.00719	.01182	.974	-.0457	.0313
	Dir.Rep	**-.09280***	**.00946**	**.000**	**-.1236**	**-.0620**
	Peer	-.00503	.00908	.982	-.0346	.0245
	Sup.Boss	**.06607***	**.01664**	**.001**	**.0119**	**.1203**
Boss	Self	.00719	.01182	.974	-.0313	.0457
	Dir.Rep	**-.08561***	**.01042**	**.000**	**-.1195**	**-.0517**
	Peer	.00216	.01007	1.000	-.0306	.0350
	Sup.Boss	**.07325***	**.01720**	**.000**	**.0172**	**.1293**
Dir.Rep	**Self**	**.09280***	**.00946**	**.000**	**.0620**	**.1236**
	Boss	**.08561***	**.01042**	**.000**	**.0517**	**.1195**
	Peer	**.08777***	**.00716**	**.000**	**.0645**	**.1111**
	Sup.Boss	**.15887***	**.01567**	**.000**	**.1078**	**.2099**
Peer	Self	.00503	.00908	.982	-.0245	.0346
	Boss	-.00216	.01007	1.000	-.0350	.0306
	Dir.Rep	**-.08777***	**.00716**	**.000**	**-.1111**	**-.0645**
	Sup.Boss	**.07109***	**.01544**	**.000**	**.0208**	**.1214**
Sup.Boss	Self	**-.06607***	**.01664**	**.001**	**-.1203**	**-.0119**
	Boss	**-.07325***	**.01720**	**.000**	**-.1293**	**-.0172**
	Dir.Rep	**-.15887***	**.01567**	**.000**	**-.2099**	**-.1078**
	Peer	**-.07109***	**.01544**	**.000**	**-.1214**	**-.0208**

***Bold indicates significant mean difference at the 0.01 level.**

Leadership Effectiveness Dimension

Group(I)	Group(J)	Mean Difference (I-J)	Std. Error	Sig.	99% Confidence Interval Lower Bound	99% Confidence Interval Upper Bound
Self	Boss	**-.31517***	**.01224**	**.000**	**-.3550**	**-.2753**
	Dir.Rep	**-.43789***	**.00980**	**.000**	**-.4698**	**-.4060**
	Peer	**-.27804***	**.00923**	**.000**	**-.3081**	**-.2480**
	Sup.Boss	**-.23873***	**.01790**	**.000**	**-.2970**	**-.1804**
Boss	**Self**	**.31517***	**.01224**	**.000**	**.2753**	**.3550**
	Dir.Rep	**-.12272***	**.01150**	**.000**	**-.1601**	**-.0853**
	Peer	**.03713***	**.01101**	**.007**	**.0013**	**.0730**
	Sup.Boss	**.07643***	**.01888**	**.001**	**.0149**	**.1379**
Dir.Rep	**Self**	**.43789***	**.00980**	**.000**	**.4060**	**.4698**
	Boss	**.12272***	**.01150**	**.000**	**.0853**	**.1601**
	Peer	**.15985***	**.00822**	**.000**	**.1331**	**.1866**
	Sup.Boss	**.19916***	**.01740**	**.000**	**.1425**	**.2558**
Peer	Self	**.27804***	**.00923**	**.000**	**.2480**	**.3081**
	Boss	**-.03713***	**.01101**	**.007**	**-.0730**	**-.0013**
	Dir.Rep	**-.15985***	**.00822**	**.000**	**-.1866**	**-.1331**
	Sup.Boss	.03930	.01708	.145	-.0164	.0950
Sup.Boss	**Self**	**.23873***	**.01790**	**.000**	**.1804**	**.2970**
	Boss	**-.07643***	**.01888**	**.001**	**-.1379**	**-.0149**
	Dir.Rep	**-.19916***	**.01740**	**.000**	**-.2558**	**-.1425**
	Peer	-.03930	.01708	.145	-.0950	.0164

***Bold indicates significant mean difference at the 0.01 level.**

Table 8 - Random Rater Response and Leniency Bias Results

Random Rater Response

Percentages of raters with $r_{wg} > .70$ in random rater response testing

Group	Analytical	Interpersonal	Courage	Leadership Effectiveness
Self	89%	95%	92%	93%
Boss	89%	95%	90%	98%
Direct Report	89%	88%	81%	98%
Peer	89%	93%	88%	98%
Supervisor's Boss	89%	94%	90%	98%

Leniency Bias

Percentages of raters with $r_{wg} > .70$ in leniency testing

Group	Analytical	Interpersonal	Courage	Leadership Effectiveness
Self	75%	84%	81%	82%
Boss	81%	84%	78%	95%
Direct Report	76%	74%	69%	94%
Peer	81%	83%	76%	93%
Supervisor's Boss	80%	85%	77%	94%

Table 9- Intraclass Correlation Coefficients by Dimension and Rater Group

Group		95% CI			F Test with True Value 0			
		ICC	Lower	Upper	Value	df1	df2	Sig*
Analytic Dimension								
Self	Single Measures	.38	.363	.389	5.816	4176	29239	.000
	Average Measures	.83	.820	.836	5.816	4176	29239	.000
Boss	Single Measures	.55	.538	.562	10.774	4411	30884	.000
	Average Measures	.91	.903	.911	10.774	4411	30884	.000
Dir.Rep	Single Measures	.56	.552	.566	11.143	14475	101332	.000
	Average Measures	.91	.908	.912	11.143	14475	101332	.000
Peer	Single Measures	.57	.566	.578	11.697	16826	117789	.000
	Average Measures	.92	.913	.916	11.697	16826	117789	.000
Sup.Boss	Single Measures	.57	.549	.588	11.529	1617	11326	.000
	Average Measures	.91	.907	.920	11.529	1617	11326	.000
Interpersonal Dimension								
Self	Single Measures	.37	.362	.386	7.571	4426	44270	.000
	Average Measures	.87	.862	.874	7.571	4426	44270	.000
Boss	Single Measures	.57	.556	.579	15.429	4698	46990	.000
	Average Measures	.94	.932	.938	15.429	4698	46990	.000
Dir.Rep	Single Measures	.61	.605	.616	18.240	15801	158020	.000
	Average Measures	.95	.944	.946	18.240	15801	158020	.000
Peer	Single Measures	.62	.615	.626	18.962	18287	182880	.000
	Average Measures	.95	.946	.948	18.962	18287	182880	.000
Sup.Boss	Single Measures	.60	.587	.622	17.808	1783	17840	.000
	Average Measures	.94	.940	.948	17.808	1783	17840	.000
Courage Dimension								
Self	Single Measures	.49	.469	.500	4.761	4454	13365	.000
	Average Measures	.79	.780	.800	4.761	4454	13365	.000
Boss	Single Measures	.54	.525	.553	5.677	4887	14664	.000
	Average Measures	.82	.816	.832	5.677	4887	14664	.000
Dir.Rep	Single Measures	.48	.471	.487	4.674	16086	48261	.000
	Average Measures	.79	.781	.791	4.674	16086	48261	.000
Peer	Single Measures	.53	.523	.537	5.506	19159	57480	.000
	Average Measures	.82	.814	.823	5.506	19159	57480	.000
Sup.Boss	Single Measures	.55	.528	.572	5.889	1876	5631	.000
	Average Measures	.83	.817	.842	5.889	1876	5631	.000
Leadership Effectiveness Dimension								
Self	Single Measures	.46	.443	.472	5.214	4384	17540	.000
	Average Measures	.81	.799	.817	5.214	4384	17540	.000
Boss	Single Measures	.77	.765	.782	18.060	4870	19484	.000
	Average Measures	.95	.942	.947	18.060	4870	19484	.000
Dir.Rep	Single Measures	.83	.823	.830	24.840	16529	66120	.000
	Average Measures	.96	.959	.961	24.840	16529	66120	.000
Peer	Single Measures	.79	.788	.796	20.050	19392	77572	.000
	Average Measures	.95	.949	.951	20.050	19392	77572	.000
Sup.Boss	Single Measures	.79	.773	.798	19.331	1901	7608	.000
	Average Measures	.95	.945	.952	19.331	1901	7608	.000

*All reported ICC's are significant for single rater and rater group in each dimension.

Vita

Kim Gower

3924 Freeport Place, #15
Richmond, VA 23233

KimGower1@gmail.com **804.335.7835**

Professional Experience

Virginia State University **Fall, 2010 – present**
Reginald F. Lewis School of Business
Assistant Professor, Management and Marketing Department

- Organizational Policy and Corporate Strategy Capstone (COBU 400): Developed and taught*
- Corporate Sustainability (COBU 304): Developed and taught*

*Courses incorporate semester long experience-based service learning projects for internal and external VSU stakeholders, all digital curriculum delivery to allow for continuous improvement and contemporary relevance, integrated modality to promote critical thinking and application of management, marketing, and finance education

Committees:

- VSU CARES Community Health Initiative: School of Business Representative, 2010-2012
- VSU Strategic Planning Task Force: Main Committee, RFP Review and Vendor, 2010 Selection, Faculty Affairs, and Graduate Studies Subcommittees
- School of Business Curriculum and Assessment Committee: Integrated, all digital curriculum using Waypoint assessment, 2010-2011
- School of Business Research Committee, 2010-2011
- Chair, School of Business Strategic Plan Revision Committee, 2011

Virginia Commonwealth University, Richmond, Virginia **Fall, 2005 to Fall, 2010**
Full Time Adjunct Instructor* – School of Business

***Service Learning Designated Course Instructor**

- *Organizational Behavior*
- *Organizational Communication*
- *Human Resource Management*
- *ESL*

VCU Graduate Assistantship **2008 to 2009**

Business Clarity LLC
Owner, Management Consulting Firm **2001 to present**

- Cross Industry and Cross Channel Experience, Profit and Not for Profit
- Specific Projects: Credit and Collection, Marketing, Team Building, Time Management, Personnel Selection and Retention, Customer Service
- Comprehensive Projects: Small Business and Professional Services Start Up and Operations

South and North Manitou LLC
Managing Partner, Interior Design Service and Retail Operations **2001 to 2005**

TeleSmart
Account Manager, Energy Industry Customer Services **1999-2001**

CMS A/R Services
Account Manager and Consultant, Energy/Telecomm Industry **1994-1999**

CareFree Building Products
Marketing Manager **1992-1994**

Gateway Systems
Marketing Manager, Application Software Development **1990-1992**

Lawyers Weekly Publications
Account Manager **1986-1990**

Publications

Journal Articles

Gower, K., & Ritter, B.A. (2010). Not a Pronoun: A Transgender's Professional Journey. *The CASE Journal, 7*(1).

Barker, R.T., & Gower, K. (2010). Strategic application of storytelling in organizations: Towards effective communication in a diverse world. *The Journal of Business Communication, 47*(3), 295-312.

Barker, R.T., & Gower, K. (2009). Use of uncertainty reduction and narrative paradigm theories in management consulting and teaching: Lessons learned *Business Communication Quarterly, 72*(3), 338-341.

Other Publications

Gower, K. (2011). Book Review: Missing Data: A Gentle Introduction; McKnight, McKnight, Sidani, Figueredo, Authors. *Organizational Research Methods*.

Manuscripts Under Review and Resubmission

Banks, G.C., Batchelor, J., Seers, A., O'Boyle, E., & Gower, K. More than just quid pro quo: A meta-analytic review of team-member exchange (TMX) theory.

Manuscripts in Preparation

Gower, K., Pollack, J.M., & Ritter, B.A. Are professors entrepreneurs? Applying the novice, serial, and portfolio entrepreneurial opportunity identification framework to the satisfaction and performance outcomes of college students.

Gower, K., Watters, S. D., & Williams, M.L. Using and interpreting the scree test in exploratory factor analysis.

Research Interests

Leadership

- Implicit Theories of Leadership
- 360/Multisource Feedback
- Leadership Competencies

Organizational Behavior

- Diversity: Gender, Cultural, Sexual Orientation
- Teams
- Learning Orientation
- Organizational Communication
- Management Education – Experience-based Learning, Service Learning

Education

Ph.D.
Expected Spring, 2012

Virginia Commonwealth University
Major: Management/Organizational Behavior
Minor: Entrepreneurship

Dissertation Topic:

Multisource feedback leadership ratings: Analyzing for measurement invariance and comparing rater group implicit leadership theories.

M.B.A.

University of Michigan – Flint; 1995

B.S.B.A.

Michigan Technological University; 1986
Major: Business Administration
Minor: Marketing

Conference Participation

Presentations

Gower, K., Watters, S. D., Williams, M.L. (2009). *Using and interpreting the scree test in exploratory factor analysis*. Annual meeting of the Southern Management Association, Asheville, NC.

Kim, N.Y. and Gower, K. (2008). *Trusting behavior and the nascent entrepreneur: The key role of trust in the critical pre-start up process.* Annual meeting of the Southern Management Association, St. Pete Beach, FL.

Gower, K. and Kim, N.Y. (2008). *Authentic leadership: Selection and development of effective organizational leaders.* Annual meeting of the Academy of Management, Anaheim, CA.

Gower, K. and Kim, N.Y. (2008). *The critical roles of trust and justice in effective organizational strategic alliances*. Annual meeting of the Academy of Management, Anaheim, CA.

Barker, R.T. and Gower, K. (2007). *Strategic application of storytelling in multi-national organizations: Towards effective cross cultural communication.* Semi-annual meetings of the International Academy of Business and Public Administration Disciplines, Orlando, FL.

Workshops

Teaching Theme Committee Member (2011). Organized and presented the Professional Development Workshop: Discussing the Undiscussables. Annual meeting of the Academy of Management, San Antonio, TX.

Watters, S. & Gower, K. (2011). Organized the Professional Development Workshop: *The Craft of Reviewing*. Annual meeting of the Academy of Management, San Antonio, TX.

Teaching Theme Committee Member (2010). Organized and presented the Professional Development Workshop: *Team Assessment.* Annual meeting of the Academy of Management, Montreal, QC.

Gower, K. (2010). Organized the Professional Development Workshop: *The Craft of Reviewing*. Annual meeting of the Academy of Management, Montreal, QC.

Gower, K. & Willis-Amat, S. (2009). Organized the Professional Development Workshop: *The Craft of Reviewing*. Annual meeting of the Academy of Management, Chicago, IL.

Teaching Theme Committee Member (2009). Organized and presented the Professional Development Workshop: *Changing the world through education as it changes around you*; *Teaching Implications in a Global Environment.* Annual meeting of the Academy of Management, Chicago, IL.

Gower, K. (2009). Organized and facilitated of a workshop entitled *Putting the Forming Back In Groups-Using a Student Developed Group Formation Project to Improve Group Performance.* Annual meeting of the Organizational Behavior Teaching Conference, Charleston, SC.

Gower, K. (2008). Facilitator in Professional Development Workshop entitled *Future of Leadership Theory*. Annual meeting of the Academy of Management, Anaheim, CA.

Sleeth, R., Gower, K., and Ritter, B. A. (2008). Organized and presented workshop entitled *An Evolving Model for the Classroom as a Project Organization.* Annual meeting of the Organizational Behavior Teaching Conference, Wellesley, MA.

Gower, K. and Sleeth, R. (2007). Organized and presented workshop entitled *The Classroom as an Organization.* Annual meeting of the Organizational Behavior Teaching Conference, Malibu, CA.

Grants and Awards

- Reaching Out MBA GLBT Case Development Research Stipend (2009) - $1500.
- Kauffman Foundation-Fellowship for Entrepreneurship Seminar (2008) - $1200.
- Phi Kappa Phi Honor Society

Service

Professional Organizations

- Teaching Theme Committee (TTC), Academy of Management (2008, 2009, 2010, 2011)

University

- VSU: School of Business Representative to VSU CARES Health Initiative, 2010-2012
- VSU: RFLSB Curriculum and Assessment Committee, Pilot Group-Waypoint Assessment Software, 2010, 2011
- VSU: RFLSB Going On Social Learning Platform Pilot Program, 2011
- VSU: Strategic Planning Task Force: Vender Selection, Faculty Affairs, and Graduate Studies Committees, 2011
- VSU: RFLSB Research Committee, 2010-2011
- VSU: Organizational Policy and Corporate Strategy capstone and Corporate Sustainability curriculum development, all digital format, 2010-2011
- VCU: President, Vice President and Co-Founder of R.A.M.S.- A graduate student organization formed to enhance management research education (2007-2010)
- VCU: Ph.D. Student Teaching Portfolio Co-Developer (Fall, 2007)

Reviewer & Discussant

- Organizational Behavior Teaching Conference Reviewer (2010)
- Ad Hoc Reviewer, Management Communication Quarterly (2009)
- Organizational Behavior Teaching Conference Reviewer (2009)
- Southern Management Association Discussant (2008)
    - Research Methods
    - Entrepreneurship
- Southern Management Association Reviewer (2008)
    - Entrepreneurship
    - Organization Theory
- Southern Management Association Reviewer (2007)
    - Entrepreneurship
    - Group and Team Building
    - Organizational Communication

Memberships

- Academy of Management
- Southern Management Association
- Organizational Behavior Teaching Society

- Center for the Advancement of Research Methods and Analysis (CARMA)
- The CASE Journal

Supplemental Training

Center for Advanced Research Methods and Analysis (CARMA)

- Live and Webcast Seminars; 2006, 2007, 2008, 2009, 2010
- Short Course Attendee; 2007, 2008, 2009, 2010

Additional Teaching Experience

Spring Arbor University, Traverse City, Michigan **2002 to 2005**

Adjunct Instructor – School of Business

Graduate Students:MAOM

- *Organizational Communication*
- *Marketing*
- *Practical Finance I*
- *Practical Finance II*

Undergraduate Students

- *Human Resource Management*

Davenport University, Traverse City, Michigan **2004 to 2005**

Adjunct Instructor – School of Business

- *Negotiation and Dispute Resolution*