5-3-2016

Social Dynamics of Leadership: Is There Value in Considering Multi-source Feedback Ratings from a Social Relations Model Perspective?

Andrea M. Bizarro
University of Connecticut, andrea.bizarro@uconn.edu

Social Dynamics of Leadership: Is There Value in Considering Multi-source Feedback Ratings from a Social Relations Model Perspective?

Andrea Marie Bizarro, Ph.D.

University of Connecticut, 2016

The present study challenges conventional use of multi-source feedback (MSF) assessment data for both leader differentiation and leader development purposes in a real organizational setting. Leadership theory was considered from an individual differences as well as a relational process perspective to investigate whether conventional MSF scoring provides adequate information regarding social processes of leadership, and whether analyzing MSF data with the Social Relations Model (SRM) provides better insights into relational processes of leadership and leader self-awareness. Application of item response theory indicated that MSF assessment items do not provide sufficient differentiation among average and above-average ability leaders. Furthermore, analysis of SRM variance components suggest that ratings of leader behavior may be subject to more substantial rating bias than has previously been estimated. Incorporating leader effectiveness outcomes in the SRM also shed light on variability in leader self-ratings and the relationship between leader self-enhancement and leader effectiveness outcomes. Results suggest that a more balanced approach to leader development may be needed in organizations to promote leader self regulation. Practical implications for organizations and directions for future research are described.

Social Dynamics of Leadership: Is There Value in Considering Multi-source Feedback Ratings from a Social Relations Model Perspective?

Andrea Marie Bizarro

B.A., Emmanuel College, **2010**

M.A., University of Connecticut, **2013**

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Social Dynamics of Leadership: Is There Value in Considering Multi-source Feedback Ratings from a Social Relations Model Perspective?

Presented by
Andrea Marie Bizarro, B.A., M.A.

Major Advisor

_____

Robert A. Henning

Associate Advisor

_____

Janet Barnes-Farrell

Associate Advisor

_____

Allan H. Church

University of Connecticut
2016

ACKNOWLEDGEMENTS

I would first and foremost like to thank my mentor and major advisor Robert Henning for his support throughout my entire graduate career. Thank you for encouraging me to pursue my ambitions and always reminding me to slow down and embrace this process.

I would like to thank my committee members - Allan Church and Janet Barnes-Farrell – for the generosity of their time and expertise, especially during the conceptualization of this study. I would also like to thank my examiners – David Kenny and Dev Dalal – for going above and beyond what was required from their roles. Your support and feedback encouraged me to persevere through many setbacks along the way.

This work would not have been possible without the support of my colleagues, my fellow graduate students, and my extended UCONN family. Your friendship and advice were always a welcome distraction from this arduous process.

To my immediate and extended family – thank you for always inspiring me with examples of your work ethic and ambition to pursue your dreams. And finally, to Garrett – I can never thank you enough for the love and encouragement you provided throughout this process.

Social Dynamics of Leadership: Is There Value in Considering Multi-source Feedback Ratings from a Social Relations Model Perspective?

Andrea Marie Bizarro, B.A., M.A.

University of Connecticut, 2016

Abstract

The present study challenges conventional use of multi-source feedback (MSF) assessment data being used for both leader differentiation and leader development purposes in a real organizational setting. Leadership theory was considered from an individual differences as well as a relational process perspective to investigate whether conventional MSF scoring provides adequate information regarding social processes of leadership and whether analyzing MSF data with the Social Relations Model (SRM) provides better insights into relational processes of leadership and leader self-awareness. Application of item response theory indicated that MSF assessment items do not provide sufficient differentiation among average and above-average ability leaders. Furthermore, analysis of SRM variance components suggest that ratings of leader behavior may be subject to more substantial rating bias than has previously been estimated. Incorporating leader effectiveness outcomes in the SRM also shed light on variability in leader self-ratings and the relationship between leader self-enhancement and leader effectiveness outcomes. Results suggest that a more balanced approach to leader development may be needed in organizations. Practical implications for organizations and directions for future research are described.

Social Dynamics of Leadership: Is There Value in Considering Multi-source Feedback Ratings from a Social Relations Model Perspective?

Organizations invest significant resources to identify, develop and retain leaders. Consequently, the concept of leadership has received a lot of attention from researchers and practitioners across all areas of organizational science (e.g., Avolio, Sosik, Jung, & Berson, 2003; Day, Fleenor, Atwater, Sturm, & McKee, 2014; Dionne, Gupta, Sotak, Shirreffs, Serban, Hao, Kim, & Yammarino, 2014; Hiller, DeChurch, Murase, & Doty, 2011). The literature is seemingly overwhelmed with various theories and development strategies all designed to improve the same bottom line: leader effectiveness. The wide availability of information is not surprising since researchers generally agree that effective leaders help an organization to maximize both social capital and employee vigor (Carmeli, Ben-Hador, Waldman, & Rupp, 2009); however, the amount of overlap in leadership theories, measurements, and methods of analysis are likely to create confusion for both practitioners and researchers.

In practice, organizations broadly use multi-trait multi-method assessments, sometimes referred to as 360 or multi-source feedback, that combine ratings of employee behavior from different rater sources primarily to identify and develop leaders that will be effective within the organization (Church, Waclawski, & Burke, 2001). Generally, these types of assessments collect ratings of leader behavior from several sources including supervisors, peers, subordinates, and sometimes customers, in order to collect as many unique perspectives on each employee as possible. These assessments also usually incorporate self-ratings as a way to identify areas where an employee's self-ratings deviate from those provided by others.

Collecting diverse perspectives on leader behavior from multiple raters and multiple sources has a long history in performance appraisal (i.e., for personnel decision-making) and in

leader development programs. Regardless of the purpose of the ratings, the general assumption is that including as many unique perspectives as possible will produce the most reliable and accurate evaluation of a leader's ability (Kornhauser, 1923) and that making leaders aware of a discrepancy between self and other scores increases leader self-awareness and motivation for development. However, these assessments may also provide an interesting lens through which to consider prevalent issues in leadership theory and research, even though they were not originally intended for this purpose. For example, several established leadership constructs appear to be redundant with each other and researchers still do not fully incorporate theories, measures, and analyses that are relevant to the leadership level of interest (Dionne et al., 2014). In an effort to address this gap, the current study investigates complementary analyses using multi-rater, multi-trait assessment data collected as part of an internal leadership program to be more in line with current thinking regarding the distinction between leaders and the process of leadership.

Although several different terms have been used in the literature to describe assessments that utilize multiple ratings (i.e., 360, multi-trait multi-method, multi-source feedback), the present study uses the general term multi-source feedback (MSF) when describing these types of assessments.

Conceptualizing Leadership

A necessary first step in the current endeavor is to identify how the term leader, or leadership, is to be conceptualized. However, as mentioned earlier leadership has been theorized and measured through many different lenses in the literature. For example, some researchers suggest that leadership is fundamentally an exchange relationship, and that a requirement of leadership is for followers to perceive the leader as someone who cares for his/her wellbeing (Yammarino & Dansereau, 2002), while others see leaders as those who possess particular

leadership skillsets (Mumford, Zaccaro, Johnson, Diana, Gilbert, & Threlfall, 2000) or personality profiles (Strang & Kuhnert, 2009). Further complicating the construct, some researchers have found that informal leaders may spontaneously emerge under certain working conditions, even in the presence of a formally identified leader or supervisor (Wheelan & Johnston, 1996). Despite a recent claim that a majority of mainstream leadership theories can all be reduced to a single factor of leader effectiveness (Rowold, Borgmann, & Diebig, 2014), the present study considers leadership theory to encompass two general perspectives: individual leader differences and relational leadership processes. Both of these theoretical perspectives of leadership are important because although individual leader differences are useful for differentiating among skills and abilities of leaders, relational processes may be more appropriate for understanding leadership in the context of dyadic exchanges or group interactions. Furthermore, relatively recent advances in leadership research support the distinction between a leader construct and a leadership construct (Avolio, Sosik, & Berson, 2013; Day, 2000) based on the understanding that each implies a unique aspect of the leadership phenomenon and what it means to be a leader.

Individual leader differences. Individual leader differences assume that leaders are individuals who posses particular traits, knowledge, abilities or skills that are necessary for effective leadership. Since this perspective largely concerns individual differences between leaders, data such as these are most appropriately considered at the individual leader level. Trait-based theories originally stem from the "Great Man" philosophy, which posits that leaders possess exceptional traits such as assertiveness, self-confidence, and cleverness (Stogdill, 1974), but early reviews of this research reported concerns over the validity of the relationship between traits and leader effectiveness outcomes (Avolio et al., 2013). Despite going out of vogue,

researchers later renewed their focus on leader traits (Kenny & Zaccaro, 1983) and found that certain personality traits had a much stronger influence on leader effectiveness than previously thought (Lord, DeVader, & Alliger, 1986). Other approaches to individual leader differentiation include comparisons of leader cognitive ability, skills, and experiences. Although cognitive ability appears to be important at all levels of leadership, political savvy and strategic ability have been shown to become more important as leaders take on more responsibility (Mumford, Campion, & Morgeson, 2007). Considering past experience, there is seemingly no clear relationship between tenure and leadership ability; however, experience with certain situations can help improve leadership (Bettin & Kennedy, 1990). Individual differences are clearly important in the study of leadership and can provide useful information, especially in the context of personnel decision-making when organizations are trying to identify potential leaders from a pool of employees or to differentiate among current leaders.

Although there is some utility in the ability to compare leaders based on traits, skills and abilities known to positively impact leader effectiveness, these inter-leader phenomena may not provide the most robust or predictive information regarding leader effectiveness. For example, as leaders take on more responsibility in an organization, training programs for certain skillsets are likely to become more ambiguous as the problems leaders face become more complex and vague. Personality traits also pose a potential problem since they are generally considered to be stable and therefore may not be the most robust measure of future leadership potential or leader development because leader personality is not likely to change over time. The best business simulations and training programs are not able to capture every subtlety of leader decisions. Therefore, current conceptualizations of leadership generally consider individual leader differences as important aspects of an individual's ability to lead effectively, but these factors are

certainly not the only determinants of effective leadership or leader development. In the present study, reference to the *leader* construct most often reflects individual difference perspectives that focus on ratings of leader traits, skills and ability.

Relational leadership perspective. Relational leadership perspectives are those that describe underlying processes of leadership, which usually involves some exchange of influence between the leader and other organizational constituents, especially followers. The relational conceptualization of leadership is primarily supported by Social Exchange Theory (Emerson, 1976), which posits that individual behavior is generally a reaction to positive or negative interactions with others. In other words, effective leaders promote mutually beneficial exchanges whereby leaders directly or indirectly elicit desired outcomes (i.e., productivity or performance) through the cultivation of personalized relationships that make others feel valued, which increases their desire to perform at a higher level.

Although Social Exchange Theory provides a valuable framework for understanding the relationship between leaders and followers, the present study considers these interactions to be a much more dynamic process of complex social tracking that involves both feedback and feedforward control. From a systems perspective, variability in performance of leaders, teams and organizations provide valuable information regarding ongoing system regulation, which has a direct impact on work outcomes such as safety, quality, and productivity (Smith, 2015). In fact, organizations can be considered high level control systems that exert influence on lower level control systems, such as those that exist within teams and even individual control systems involved in goal-setting (Edwards, 1992; Manz, 1986). For example, leaders who are able to successfully track and respond to feedback based on the performance of their followers may be more effective at managing workflow, which is likely to benefit followers as well as the quality

of their work output. At a more macro level, leaders who are able to successfully track naturally occurring organizational performance variability over time may be better positioned to exert the feedforward control that is necessary for long-term planning and shaping the future vision of the organization. The idea that compensatory social tracking is a fundamental aspect of established, dynamic social systems such as workgroups and organizations is not new (Smith & Smith, 1987), and the present study argues that social tracking processes are a critical aspect of effective leadership. Therefore, although the relational leadership perspectives described below are formally grounded in Social Exchange Theory, I propose that these exchanges are all supported by a leader's fundamental ability to support effective feedback and feedforward control of work processes and social systems rather than limited to an attitude-based exchange relationship.

Researchers contend that leadership theory has naturally been progressing from a focus on the individual to a focus on dyadic processes between leaders and other organizational stakeholders (Yammarino & Dansereau, 2002). Several of the most prominent leadership theories explicitly describe interpersonal dynamics between the leader and other organizational constituents including leader-member exchange (Graen & Uhl-Bien, 1995), transformational leadership (Bass, 1985), individualized leadership (Yammarino & Dansereau, 2002), and authentic leadership (Avolio & Gardner, 2005). The other common factor among these relational perspectives is that the study of leadership processes is inherently interpersonal and therefore most appropriately measured at the dyadic level of analysis. As leadership continues to become more nuanced, theories explaining the dynamic processes that underlie leadership become increasingly important because they have been linked to positive outcomes such as improved social capital (Carmeli et al., 2009) as well as improvements in perceived team effectiveness and productivity (Burke, Stagl, Klein, Goodwin, Salas, & Halpin, 2006). Furthermore, a relational

perspective may be better suited to understanding factors that contribute to effective leadership in complex and vague situations in ways that inter-leader perspectives are not able to. In the present study, reference to the leadership construct most often reflects these dynamic processes involved in interpersonal interactions.

Implications of Differing Leadership Perspectives for MSF

The present study argues that conventional use of MSF assessments in organizations is more aligned with individual difference perspectives as opposed to relational perspectives, and explores the value of utilizing different analysis techniques that are more closely aligned to the relational perspectives of leadership. Insights gained by these techniques may provide leaders and organizations with unique and potentially more useful information, especially in the context of leadership development. In fact, MSF research documents increasing concern with aspects of MSF other than the ratings themselves including the consideration of qualitative experiences of MSF participants as well as the coaches who provide feedback (Harrington, 2012), and there is renewed criticism regarding the conventional method of aggregating MSF scores within rater groups (Markham, Smith, Markham, & Braekkan, 2014). Some critics also question the influence of self-rating bias and therefore the validity of self-awareness measures that are based solely on rating discrepancy. The recent advancements in the current thinking among leadership and MSF assessment researchers may have implications for how MSF assessment data are routinely being used in organizations.

Multi-source feedback assessments evolved around a general need for organizations to gather multiple perspectives on leader behavior, but the application and impact of interpreting ratings across various applications of MSF data vary. Multi-source feedback assessments have two primary uses in organizations. The first application of MSF assessments is concerned with

differentiation. Organizations need a way to differentiate among leaders to determine which employees are likely to excel in leadership roles and which may need more training or coaching. A need for this type of assessment began as the economy started to shift in the early 1900s to service-oriented occupations where employee knowledge, skills and abilities were more difficult to observe and thus more difficult to differentiate (Hedge, Borman, Birkeland, 2001). An early review of psychology in the workplace notes that in order to improve reliability of employee ratings, some researchers had begun to include peer ratings to supplement the more conventional supervisor ratings of employee performance (Link, 1920). Yet despite the increase in reliability that results from combining multiple ratings of employee performance, even this early review by Link (1920) acknowledged the potential bias introduced by raters who may not be able to objectively assess performance. Nevertheless, aggregated responses from MSF assessments continue to be used as a way to differentiate among leaders and to compare leader scores on various attributes to internal or external norms.

Despite concerns raised regarding rating bias and the possible influence of rater motivations, researchers continue to recommend that adding more raters as well as different types of raters will increase the reliability of the ratings in addition to insights gained from the various perspectives of raters who know the employee in different contexts. From a classical test theory perspective, variation among raters within rater groups is often attributed to measurement error, and increasing reliability of the score by adding more ratings within each group reduces the influence of this potential bias (Scullen, 1997). However, increased reliability of ratings does not make ratings more valid predictors of leader ability. For example, in the case of performance ratings the relationship between subjective ratings of performance and objectively measured performance is relatively weak (Mabe & West, 1982; Murphy, 2008), and more modern types of

employee ratings acknowledge potential bias and make some attempt to correct for it. Researchers suggest that holding raters accountable for rating accuracy by linking MSF ratings to organizational reward systems, such as promotion decisions and bonus structures, would encourage raters to provide more accurate ratings (Bracken, 1997; Tornow, 1998). London (2001) also notes that ratings may have less bias if raters are formally trained in behavioral observation. These findings support the notion that although differentiation among leaders is necessary to reward behavior and place leaders into appropriate roles, the conventional use of aggregated MSF ratings to provide this differentiation is not without limitations from either the individual leader or relational perspectives of leadership.

The other primary application of MSF assessments is for developmental purposes. As mentioned earlier, researchers distinguish between *leader* development and *leadership* development but this distinction is a relatively recent addition to the leadership literature (Avolio et al., 2013; Day, 2000). Early leader development groups were heavily influenced by the work of Kurt Lewin in the post World War II era and are most representative of Day's (2000) conceptualization of *leadership* development. During a time period considered as a social revolution, researchers introduced development programs called "sensitivity training" to organizations (Highhouse, 2002). The goal of these sensitivity training sessions, also known as the T-Group, was to enlighten managers through group feedback sessions as to how their behavior influences larger social systems and group processes. Methods used in T-groups and sensitivity training sessions were based on democratic teachings and awareness training with the idea that leader behavior has a great influence over social systems through the direct influence of the leader's behavior on the social environment.

What started as a program to educate community and social leaders to understand increasingly diverse societies eventually evolved into early leader development programs (Slater & Coyle, 2014). A hallmark of these leader development sessions was the focus on leader self-awareness through direct confrontation and role-play with others in the session. Furthermore, researchers quickly noted that leaders were highly interested in receiving expert feedback regarding how their behavior impacted others (Bradford, 1976). In their early form, these development programs were primarily focused on group dynamics and shaping leader behavior through an understanding of how his or her actions impact the behavior of others in the immediate group. This earlier application is much more in line with the dyadic process implied by relational perspectives of leadership in contrast to modern MSF assessments that use aggregated ratings of leader ability to motivate behavior change by highlighting discrepancies between self-ratings and the ratings provided by others.

Researchers regularly debate whether or not MSF assessments should be used for both assessment and development purposes rather than serving exclusively for development purposes (London, 2001). This tension is further complicated by how expensive and time-consuming a MSF process is for organizations. Organizations are likely to be motivated to use MSF for dual purposes in order to maximize the return on their investment in the process. However, as mentioned previously, utilizing the same data to inform individual leader differentiation as well as dynamic leadership processes may prove to be detrimental to the broader organizational system since aggregated MSF scores may not provide leaders with enough feedback information on how their leadership style influences social dynamics.

Leader self-regulation. As stated previously, the present study considers relational perspectives of leadership from a control theory perspective such that a leader's ability to engage

in successful social tracking progressively improves his/her ability to exert both the feedback and feedforward control that is necessary for effective leadership. Therefore, a majority of the hypotheses proposed in the current study are concerned with the leadership development application of MSF assessment data and whether conventional applications and analyses of MSF assessment data are relevant for leader self-regulation. Leader self-regulation concerns the process through which leaders are motivated to change their behavior or engage in development activities based on feedback gathered from a MSF assessment. Control theory posits that leaders are motivated to regulate their behavior in such a way as to reduce discrepancy between their own behavior and an environmental behavior standard (i.e., other-ratings in MSF assessment; Carver & Scheier, 1982). Indeed, Ashford and Tsui (1991) demonstrated this phenomenon when they found that actively seeking critical feedback had a strong impact on perceived managerial effectiveness. Additionally, researchers find that control theory principles influence goal-setting behaviors (Campion & Lord, 1982) as well as employee stress appraisal and coping responses (Edwards, 1992). From a control theory perspective, leader self-regulation involves the ability of leaders to utilize relational feedback from social tracking or by comparing personal behaviors to established norms.

When MSF assessments are implemented as part of regular leader development programs, the performance standards set by the assessment are likely to have a powerful influence on leaders and how they view their own behavior in comparison to these standards. Therefore, application of MSF assessments within organizational systems is likely to benefit from consideration of control dynamics consistent with principles of control theory. However, there is little evidence to support the notion that MSF assessments were developed with these

types of self-regulation in mind, especially in the context of providing leaders feedback on relational processes of leadership that may be highly relevant to leader effectiveness outcomes.

Existing MSF research suggests that agreement between self- and other-ratings is an indication of leader self-awareness, and finds a positive association between leader self-awareness and leader effectiveness (Church, 1997; Atwater & Yammarino, 1997). Yet, several studies provide evidence that the discrepancy between self and other ratings is not a consistent predictor of manager goal setting (Morgan, Cannan, & Cullinane, 2005), behavior change (Smither, London, & Reilly, 2005), or current performance ratings (Atwater, Waldman, & Brett, 2002), which weakens the validity of MSF for influencing or even measuring employee behavior adjustments over time. Therefore, the current literature provides little evidence to inform whether MSF assessments inform or support leader self-regulation activities that may be extremely important for leadership development of relational processes.

Despite early research on the validity of self-awareness researchers are still unclear on what self-awareness is and whether discrepancy scores accurately reflect self-awareness in social interaction or self-awareness of individual ability (Kulas & Finklestein, 2007). In fact, some researchers have found that rating discrepancy is negatively related to self-monitoring behaviors in leaders (Church, 1997; Fletcher & Baldry, 2000). In other words, there is a research need to investigate self-awareness in MSF and whether agreement between self and other ratings is an indication that the leader is being realistic about where he/she stands on a trait, or whether the leader is simply savvy enough to know how others would rate him/her. This distinction is important for understanding leader development from a self-regulation perspective and may be relevant to inconsistencies in the current literature regarding outcomes of leader participation in MSF development programs.

Self-regulation is important for leader development activities because it is the process that motivates individual leaders to change their behaviors in line with development goals (Kanfer, 1990; Porath & Bateman, 2006). In terms of MSF assessment data, leaders presumably set developmental goals in order to address discrepancy between the leader's scores and internal or external leader norms. Additionally, comparisons of leader self-ratings to ratings provided by others can also provide incentive for leaders to address areas where there may be a gap in perspectives. Researchers acknowledge that active feedback seeking is an important aspect of leader self-regulation (Ashford & Tsui, 1991); therefore, MSF assessments that are designed to provide this information should be considered carefully especially in how feedback is communicated. As mentioned previously, raters are inherently biased in their perceptions of the leader's behavior. Therefore, a focus on discrepancy between a leader's self-ratings and aggregated ratings from others may provide overly simplistic and/or inherently misleading feedback to the leader. Furthermore, MSF assessments primarily measure leader skills and traits, which, as mentioned previously, are more suited to individual leader development as opposed to leader*ship* development. This individual leader perspective may not fully inform leadership development plans in terms of leader self-regulation in the face of dynamic relational processes associated with effective leadership as identified by LMX, transformational leadership, or individualized leadership.

Research suggests that self-regulation is an important process for leader effectiveness in day-to-day activities (Porath & Bateman, 2006) as well as during more challenging organizational activities (Taylor-Bianco & Schermerhorn, 2006). Furthermore, others in the organization, especially followers, often have the expectation that leaders should be able to adapt to changing situations, which requires a high level of self-regulation (Sosik, Potosky, & Jung,

2002). Therefore, cultivating self-regulation among a leader population should be a high priority for an organization. The alternative relational models used in this study may serve as a way for organizations to explore different ways of assessing self-regulation in their leaders with existing MSF assessment data. The research questions and hypotheses below are all directed to understanding: (1) how MSF assessment data informs leadership programs in terms of individual leader and relational processes from a psychometric perspective, (2) whether alternative scoring of MSF assessment data provide more relevant information regarding relational processes in leadership, and (3) whether alternative scoring of MSF assessment data can better inform leadership development plans by calling attention to issues relevant to leader self-regulation.

General Overview of Present Study

Part I of the present study will examine the impact that long-term use of MSF has, for both development and decision-making purposes within a single working population, on the psychometric properties of MSF ratings. The main research question in Part I is whether information provided by the test changes as a result of continued use of the same assessment for multiple purposes over several years. The overall goals of Part I are to determine if and how scores change over time, whether scores on leadership competencies improve over time, and whether changes in scores over time are due to leader development (i.e., changes in leader behavior) or represent psychometric artifacts that result from changes in assessment practices over time within the same general leader population. Item response theory (IRT) will be used to investigate parameter drift in MSF items and test scores over time.

Part II of the study will investigate whether group dynamics among raters influence leader MSF scores and leader performance. The main goals in Part II of the study are to understand how analyzing MSF data at the dyadic and group level using the SRM model predicts

leader competency scores and whether these measures add unique information beyond that of conventional agreement or aggregate scores. Leaders participating in MSF assessments in large organizations are likely asked to rate other leaders participating in the same assessments as either a peer, subordinate, or supervisor. The availability of these data provides the opportunity to incorporate round-robin scoring methods such as those used in the Social Relations Model (SRM; Kenny, 1994), which is a component analysis technique commonly used to study small group dynamics but which I propose to adapt for use here. Grounded in G Theory, this analysis can provide insights into relational dynamics that may be driving ratings of leader behavior.

Finally, Part III of the dissertation expands on dyadic analyses from the SRM to estimate leader self-enhancement in a new way. Leader self-enhancement is a measure of inflated sense of self as evidenced by self-ratings compared to ratings provided by others as well as the ratings the leader provides for others. The main goal in Part III of the present study is to determine whether using the SRM to calculate leader self-enhancement is a better predictor of leader effectiveness outcomes, such as turnover and leader quality, than conventional discrepancy scores. Self-enhancement assessed with the SRM can factor in two important aspects of self-perception in terms of relational leadership processes: (1) whether the leader rates him/her self higher or lower than other raters and (2) whether his/her self-ratings differ from the way he/she rates others.

Multi-Source Feedback Assessment and Dataset Characteristics

The dissertation will use archival MSF data collected annually over a period of approximately 5 years within a single organization to accomplish the goals described above. These data were collected as part of a large consumer goods organization’s systematic leader assessment and development program.

Leadership competency model. The host organization's leader competency model is important to the present study because the MSF assessment items are tailored to dimensions and leadership tiers outlined in the model. The leadership competency model describes ideal employee behaviors related to two dimensions relevant to the present study: (1) task-related skills and (2) socio-emotional skills. The behaviors described in each dimension are further specified as the three tiers of the leadership competency model. The *foundational tier* of the leader competency model describes leadership behaviors for each dimension that all employees should strive to achieve such as active listening and understanding of business functions. Leaders participating in MSF assessments tailored to the *foundational tier* are typically first time or recently promoted managers who are learning how to work with subordinates. The *middle tier* of the leader competency model describes leader behaviors that experienced middle managers should adhere to such as collaborating across teams and developing expertise in specific parts of the business. These leaders are not only more experienced than *foundational tier* leaders; they generally have greater responsibility than *foundational tier* leaders. Finally, the *top tier* of the leader competency model describes behaviors the senior leaders of the organization should adhere to. These leaders have the most responsibility in the organization and their behaviors are more abstract such as setting a compelling vision and supporting an overall business strategy.

Each successive tier of the leader competency model builds on the behaviors outlined in the previous tier. This way, even leaders in the *top tier* are still held accountable for basic functional business knowledge and skill development that *foundational tier* leaders are accountable for. Each of the three leadership tiers includes MSF assessment items that are relevant to the three dimensions described above. The present study refers to the three MSF assessments generally with the understanding that each assessment encompass the respective

leadership tier the assessment is tailored for. In some cases, hypotheses are written with considerations that are specific to a particular leadership tier. A majority of hypotheses focus on the *middle tier* of leadership because this tier tends to be the most stable over time. Leaders spend most of their career as *middle tier* leaders so this group is likely to have the most consistent group of leaders over time.

The MSF assessment. The MSF assessment was customized for the organization for two purposes: leader development as well as one of several inputs used for personnel decision-making. In order to assist leaders in improving their behavior, the MSF assessment asks peers, supervisors, and subordinates to provide ratings of leader behaviors as they relate to the organization's leadership competency model. The competency model identifies traits and behaviors that the ideal leader should possess and the assessment measures two dimensions of leadership related to the hypotheses in the present study: (1) task-related skills and (2) socio-emotional skills.

To further support the organization's leadership model, three different versions of the MSF assessment were created to accommodate leaders at different levels of the organization. All employees are expected to demonstrate certain behaviors related to the three leadership dimensions mentioned above, which are reflected in the core content included on all three MSF assessments of the following groups: *foundational tier*, *middle tier*, and *top tier*. The assessments for *middle tier* and *top tier* leaders also include items specific to the behaviors expected at their respective levels of the organization in addition to the core items from the more general *foundational tier* assessment.

Dataset characteristics. Using data from a single organization collected over multiple years provides unique opportunities to explore previously unanswered questions regarding the

potential effects of repeated administration of MSF assessments over a long time period as well as to apply scoring alternatives that incorporate round-robin data. First, continued use of the same pool of items provides robust data that can be used in more rigorous psychometric analyses than have been previously explored; such as utilizing item-response theory (IRT) to understand how item discrimination may change over time and between participant groups. Additionally, the broad use of MSF assessments within the same leader population increases the likelihood that each leader in the population has both received feedback and provided feedback in various administrations of the assessment. This final point allows for the incorporation of each leader's response patterns in addition to those provided by the raters, providing valuable indications of leader behavior beyond those provided by the rating sources alone. Other advantages of utilizing this comprehensive longitudinal dataset are explained in more detail later on when specific analyses are described.

Part I: Psychometric Analyses

It is possible that continued use of MSF assessments within the same employee population might result in the degradation of the psychometric properties of test items and scores over time, with differential effects occurring across subgroups of the population. The data used in the present study provides an opportunity to understand how the same items operate within assessments administered at distinct levels of the organization and to look at how tests within each level perform after several years of repeated administration. This first section of the present study investigates whether the three distinct assessments lose some of their integrity over time and whether items tailored specifically for each type of leader assessment provide better differentiation among MSF ratings of leaders. Item differentiation is especially important for leader development given that leaders rely on feedback from these items to help manage leader

self-regulation in relation to behavior change initiatives such as receiving additional training or being involved in more formal development programs.

As mentioned previously, MSF assessments in the present organization contain two item sets that each asks raters to provide feedback on leader ability regarding the leadership dimensions (i.e., task-related skills and interpersonal skill) and tailored to the leadership tiers. The first is a set of 23 core items that are the same for all leaders. These items are most relevant to the fundamental behaviors expected of *foundational tier* leaders within each leadership dimension, and are considered the foundation of the organization's broader leader competency model. The MSF assessment for *foundational tier* leaders only includes these 23 core items. The second set of items appears on the *middle* and *top tier* leader MSF assessments, and is specifically tailored to address the leadership behaviors associated with each dimensions of the leader competency model relevant to *middle* and *top tier* leaders. In general, leaders use item-level scores to inform their personal development program by focusing on specific behaviors relevant to their lowest scoring items or items with the greatest discrepancy between self and other ratings. Aggregated test scores are also used as inputs for performance appraisal and organizational decision-making. Each dimension of the MSF assessment is scored for each manager, and these individual manager scores are usually compared to an overall organization average, or internal benchmark score. The contexts in which MSF assessments are conducted have implications for both rater and participant motivation, which ultimately influences how raters respond to MSF items (London, 2001); therefore, Part I of the present study considers both the item-level and dimension-level psychometric properties of the organization's MSF assessment.

Earlier research suggests that formal communication regarding the purpose of the MSF assessment is crucial for successful implementation (Brutus & Derayah, 2002); however, these researchers generally did not investigate whether the same MSF assessments are valid for use in both decision-making and development programs rather than in development programs alone. Raters participating in confidential MSF assessments that are only used for leader development purposes may be more honest in their responses than those who provide ratings for assessments that will be used for decision-making purposes such as succession planning and performance appraisals (Tornow, 1998). Presumably, raters may be more honest and willing to provide lower ratings in a development context because low ratings would not have an adverse impact on the rated leader's standing in the organization. Researchers have also argued that using an MSF assessment for decision-making is counterproductive to behavior change goals for development (Dalton, 1997). Furthermore, the process of providing ratings on behavior for development is fundamentally different from the process for ratings focused on judgment and evaluation, especially for supervisors (London & Tornow, 1998). Organizations that do not regularly monitor the psychometric properties of assessments over time risk potential shifts in how well the assessment is able to predict leader performance over time. Therefore, it becomes much more important to monitor the psychometric integrity of an assessment used within the same organization over time for both decision-making and development purposes.

As the popularity of MSF assessments started to grow, more robust psychometric analyses, such as item-response theory (IRT), were introduced to the literature. In response to early criticism that MSF ratings violate the independent observation assumptions of IRT, Craig and Kaiser (2003) demonstrated that IRT results from MSF ratings did not result in undue bias. In other words, IRT results were the same whether or not the researchers accounted for nesting

of ratings within leaders and the possibility that the same raters rate multiple individuals. Craig and Kaiser (2003) used a relational perspective to argue that even though ratings of individual leaders are nested, and some raters rate multiple leaders, errors are less likely to be correlated because each leader presumably has a completely unique relationship to each rater. Item-response theory analyses of MSF data to date have focused on the measurement equivalence of ratings across rater groups (Facteau & Craig, 2001; Penny, 2003) as well as various demographic and contextual differences including gender and environmental complexity (Penny 2010) as well as ratings for leaders categorized as over estimating or underestimating ability (Kulas & Finklestein, 2007). However, to the author's knowledge, researchers have yet to apply item-response theory to detect parameter drift as a result of the use of MSF assessments over time.

It stands to reason that, if MSF assessments are used repeatedly in leader development efforts, then leaders will eventually conform to the performance standards measured by the leadership competencies as defined by these assessments, especially if existing leaders regularly exhibit the desired skilled behaviors (Kempster & Parry, 2014). The learning that occurs as part of leadership development could, in turn, influence the functioning of the MSF assessment in a similar way that long-term use of other standardized tests do. Therefore, long-term use of MSF could impact the integrity of MSF ratings, long-term individual employee development, and the broader organizational culture, all of which have implications regarding the usefulness of this data to guide leader self-regulation processes. Therefore, the first part of this evaluation effort examines the use of MSF assessments over time, both across and within leader ratings.

Parameter Drift

Item-response theory research on test development has shown that individual items as well as entire test forms can change over time; a phenomenon commonly referred to in the

literature as parameter drift (Goldstein, 1983). Environmental or cultural shifts can contribute to this problem when the meaning of constructs or definitions of relevant terms change over time, which in turn can change the parameter estimates for individual items or even entire tests. Early demonstrations of parameter drift show the impact that parameter drift has on test taking over time, and also how changes in item and test functioning are sometimes misinterpreted as changes in respondent ability (Goldstein, 1983; Bock, Muraki, & Pfeiffenberger, 1988). Common examples of parameter drift can be seen in educational testing when tests fail to change with the curriculum. In other words, if test content remains the same, but less attention is paid to some content in the curriculum, the difficulty of items related to this content will increase over time as students become less familiar with it. This disconnect can result in changes in test scores over time that have less to do with student ability and more to do with changes in current educational priorities that are no longer aligned with the test.

In the case of MSF assessments in organizations, changes in leadership or organizational transformations can impact the relevance of particular items on MSF assessments in the same way that educational tests can be impacted. There are two main ways in which shifts in test properties could have occurred with the MSF data used in the present study. First, the tiered assessments tailored to the leader's level in the organization can be considered as similar to the curriculum shift example in educational testing (Bock et al., 1988). Changes in descriptions of behaviors in the leadership competency model at higher levels of the organization may impact how leaders and raters respond to the same items because an organizational shift in leader focus may change the relative importance or significance of various competencies measured on MSF assessments. Furthermore, certain skillsets may become more relevant for leaders as they gain more responsibility (Mumford, et al., 2007). For example, raters may be more focused on leader

task-related skills compared to socio-emotional skills especially when providing ratings for *senior leaders* because the organization places greater responsibility on these high-level leaders to drive business results.

The second main way in which test properties could change over time is through slight shifts in communication regarding assessment feedback, especially for leaders who participate in multiple cycles of MSF testing. In the target organization, leaders are oftentimes provided one-on-one coaching feedback after receiving MSF ratings. These coaching sessions will likely change how the leader observes his and others' behaviors based on what he learns in these feedback sessions. This could increase the amount of informal behavior appraisal beyond what would naturally occur within the organization, which could then impact how well the MSF assessment accurately differentiates among leaders.

Parameter drift can be measured at the item, scale, or test level and, when used for assessments within an organization, should be examined at the same level personnel decisions are made (Drasgow & Hulin, 1990). The test or competency would demonstrate parameter drift if variations in ratings for individuals assumed to be at similar standings on the trait differ not only as a function of each leader's standing on the trait, but also as a result of time. The impact of time on item responses is considered in two different contexts in the present study. First, responses could vary as a function of the year the assessments were administered. Item discrimination in general may change as leaders and raters become more familiar with the assessment content over time. Second, individual leader scores are likely to change over time, which can either be attributed to leader development, or to changes in item discrimination across assessments. Within-leader analyses of score changes over time will provide some insight as to

whether these changes can be attributed to leader development efforts or simply to a change in item functioning over time.

Hypotheses. Part I of the dissertation examines how core content compares to unique content in discriminating leader ability and how repeated administrations might impact test discrimination. Item-level and dimension-level comparisons will be conducted separately. First, psychometric properties of the items included on the tiered MSF assessments (i.e., *foundational tier*, *middle tier*, and *top tier* leaders) are hypothesized to change as a result of the tier of the leader being rated. These changes are hypothesized to have implications for item-level properties of the assessments. Item-level analyses might be more relevant to the MSF data used for leader development because leaders often focus on development activities related to their lowest scoring assessment items.

IRT-H1: The 23 core items will demonstrate lower discrimination for (a) middle tier *leaders and for (b)* top tier *leaders than for* foundational tier *leaders.*

IRT-H2: Items written specifically for middle tier *or* top tier *leader assessments will provide better discrimination than core items for (a)* middle tier *and (b)* top tier *leaders.*

Dimension-level analyses are hypothesized to be most relevant to the MSF data used as inputs for performance appraisal because these aggregate scores are used to determine how well leaders perform on each dimension compared to others. Therefore, psychometric properties will also be measured at the test level in order to understand how the test as a whole is able to discriminate leaders. As stated previously, continued use of the same assessment over time within the same organization is likely to have an influence on how raters observe behavior. If leaders do in fact gradually shift to more frequent and informal observations of behavior over time, then leader behavior adjustments may be less obvious to raters, which could reflect in

lower discrimination of the test as a whole among leaders over time. Furthermore, parameter drift assessments can provide some insight if changes in leader scores over time can be attributed to time or to some other factor such as behavioral adjustments on the part of the leader. The following hypotheses focus on leaders who have participated in the *middle tier* assessments since this group of leaders is the most stable and consistent over multiple years, which is ideal for longitudinal analyses.

IRT-H3: Test discrimination will decrease over a period of five years such that test information is higher in 2009 than in 2014 for middle tier *assessments.*

IRT-H4: Time will predict a significant shift in test parameters over a period of 5 years for middle tier *leaders who have participated in MSF assessments more than once from 2009 to 2014.*

Part II: Exploring Novel Methods in Multi-Source Feedback Non-conventional Approaches

As stated previously, some researchers argue that MSF data should be used for both development and decision-making purposes because the organization dedicates significant resources to collect these data, and to use these data for only one purpose could be considered wasteful (London, 2001). However, use of MSF for personnel decision-making employs an inherently different level of theory and analysis than use of MSF for leadership development, and furthermore, conventional one-sided discrepancy scores provide little information relevant to leader self-regulation. In the section that follows, the present study considers the ways that MSF data has conventionally been analyzed and used in the past and then introduces a more novel analysis approach which could potentially provide a supplemental way to analyze MSF data in a way that would be more relevant to relational perspectives of leadership and leader self-

regulation, which are inherently more relevant to personnel decision-making, as opposed to the individual leader perspective of ability scores.

Conventional Analysis Methods of MSF

Assumed advantages of collecting data from multiple raters with MSF assessments are that each rater provides a piece of unique information on leader behaviors and that having multiple ratings increases the reliability of ratings. As stated previously, ratings are then combined to provide overall scores across or within rater groups (i.e., subordinates, peers, etc.). The practice of aggregating MSF ratings across and within rater groups is regularly debated in the literature and has recently come under renewed scrutiny because conventional MSF aggregation methods do not consider within-rater agreement before combining MSF ratings (Markham et al., 2014). Additionally, some researchers find that peers may provide more reliable data than supervisors (Conway & Huffcutt, 1997; Viswesvaran, Ones, & Schmidt, 1996), yet other findings suggest that this increase in reliability is likely due to the fact that the peer ratings are averaged across raters compared to the sole supervisor rating (Murphy et al., 2001). Although a common practice in applied settings, aggregating responses could potentially obscure important variability among ratings that may provide unique and relevant information needed for understanding leader self-regulation.

Another popular way to analyze MSF data is to measure agreement, or variability, within and between rating groups to understand whether variance in ratings is due to idiosyncratic rater effects. Even though some MSF reports may provide leaders with an indication of the spread of ratings (i.e., highest and lowest rating compared to the mean score), high variability is most often considered a source of error as opposed to a metric used to interpret data. And furthermore, aggregated ratings alone may not be suitable to understand potential underlying rating bias or

patterns that may exist among raters. This concept is interesting because attributing variability in ratings as partially due to error suggests the possibility that discrepancy between self and other ratings are due to chance, which would undermine part of the MSF development process (Kulas & Finklestein, 2007). More recently, researchers have also discussed levels of analysis issues with conventional scoring methods (Markham et al., 2014), and proposed alternative ways to analyze MSF data (Yammarino, 2003); however, existing research analysis approaches have yet to leverage the full range of data available, including leader self-ratings, and any ratings the leader may also be providing for his/her raters.

The present study is not the first to observe that dyadic influence between leaders and other organizational constituents is a level of analysis commonly overlooked in organizational research (Gooty & Yammarino, 2011) or to suggest that conventional aggregation methods in MSF may not be the ideal analysis technique. Findings from studies analyzing variance components for MSF data reveal that researchers have long been aware of the large amount of influence raters have over leader scores and that ratings that can vary depending on the number of raters polled (Greguras & Robie, 1998) and on each rater's unique relationship to the participating manager (Church, 1997). Two methods that focus specifically on variability among MSF ratings have already been explored in the literature: within- and between analyses (WABA) and Generalizability (G) Theory. Each of these methods provides complementary information to aggregated scores and offers unique information that can be used by leaders and practitioners alike.

The use of WABA in MSF analyses is significant because the approach is designed to assess the best way to analyze data from several angles: (1) appropriateness of aggregating within rater groups, (2) analysis of agreement within self-other dyads (i.e., self-manager, self-

peer, etc.), and (3) analysis of agreement within other-other dyads (i.e., manager-peer, peer-subordinate, etc.). This method of analysis is important because it requires a systematic analysis of variance in order to determine what the most appropriate level of analysis is for the data (Dansereau, Cho, & Yammarino, 2006). For example, if analyses in the first stage of WABA indicate a low level of agreement within each rater group, then scores would not be aggregated to the rater group level. Rather, data would be analyzed at the dyadic level. Insights from each level of analysis can then be summarized and integrated as part of the feedback provided to managers for leadership development. Although the use of WABA in MSF analyses provides leaders with more information and addresses potential issues with idiosyncratic ratings, this method still overlooks the potentially important information regarding two-way dyadic ratings between the leader and others (i.e., others' ratings of the leader and the leaders' ratings of others). Additionally, this method may not be desirable in applied settings because with low agreement in Step 1, some, if not all, leaders would not have the competency scores needed to conduct comparisons with other leaders.

Yet another line of research aimed at understanding MSF ratings utilizes G Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to quantify variance attributed to various aspects of MSF assessments by directly measuring variance due to rater effects, test effects, or manager trait effects. Previous research has used G Theory in tests of construct validity to demonstrate that variation in ratings among leaders is primarily due to trait effects (i.e., individual differences between leaders in knowledge, skills and abilities) as opposed to effects due to rater type (i.e., whether ratings were provided by a subordinate, peer, supervisor, etc.) or to the test items (Conway & Huffcutt, 1997). For aggregated rater scores that truly provide a

valid measure of leader behavior, a G Theory study would observe a high level of variance attributable to leader trait and non-significant variance attributable to rater or item effects.

In recent G Theory studies however, rater effects have been shown to be relatively high across several studies in comparison to trait or source effects (Yammarino, 2003). Researchers also find that leader ratings can be attributed to individual rater effects that are not related to rater type above and beyond variance due to leader behaviors or trait effects (Mount, Judge, Scullen, Sytsma, & Hezlett, 1998). In other words, variance in ratings seems to be a function of each individual's unique perspective on the manager and this variation has little to do with whether the rater is a supervisor, subordinate, or peer. The strong influence that individual raters have been shown to have on MSF ratings is yet another indication that complex social relationships are likely to play a role in how others individually appraise and rate the leader's behavior. Researchers may be tempted to credit these variations to error variance; however these findings could also be interpreted as providing additional support for the relational process perspectives of leadership mentioned previously. More specifically, these research findings support the notion that simple aggregation of ratings may not provide the best insights into leader self-regulation behavior, especially behavior related to social dynamics, and that researchers may need to consider dyadic effects between leaders and raters within rater groups.

A Relational Perspective

An individual approach to leadership assessment is somewhat necessary if the goal is to identify and differentiate among leaders at the individual level. However, as mentioned previously, relational perspectives suggest that effective leadership is also a social exchange process between leaders and followers. For example, studies of transformational leadership (Bass & Avolio, 1994) suggest that the best leaders are able to inspire and motivate their followers to

perform through the leader's charisma and cultivation of high quality relationships. These leaders are also more likely to actively self-regulate their behavior and adjust their leadership style to align with the needs of each social interaction. Conventional MSF aggregation methods offer little insight into these unique soft skills, especially leader self-regulation, because each rater is likely to have a somewhat unique relationship with the leader. In other words, raters cannot be considered interchangeable agents that are part of a rating instrument, as is the case with conventional assumptions of MSF, because they each have a unique perspective on leader behavior based on present and past personal interactions with each leader (Murphy et al, 2001). Researchers focusing on leader-member exchange find further support for a relational perspective of leadership. For example, team members who perceive low-quality exchanges with a leader also report lower task efficacy and higher conflict (Boies & Howell, 2006). Previous research findings also illustrate how interpersonal relationships between leaders and others impact individual and team-level outcomes such as improving communication satisfaction (Mueller & Lee, 2002) and team climate (Tse, Dasborough, & Ashkanasy, 2008). Studies such as these suggest that the most effective leaders are those who cultivate productive relationships with team members by empowering each and every group member and contributing to his/her development.

Considering the above information, it is not surprising that aggregated MSF ratings are not consistently related to leader effectiveness outcomes. As mentioned earlier, it is possible that even ratings of a single leader can be considered independent because these ratings are based on a completely unique dyadic relationship (Craig & Kaiser, 2003). In fact, patterns of differentiation can be seen in MSF assessment ratings and, even within rating group types, raters tend to have relatively low agreement among ratings of leader behavior (Mount et al., 1998).

Aggregating MSF scores as a scoring approach ignores potentially important differentiation among ratings, when such differentiation could provide insights into dyadic influences of individual relationships between the participating leader and each rater as well as insights into leader self-regulation.

In addition to the unique relationships raters are likely able to cultivate with high quality leaders, raters are also likely biased by their own personal goals and motivations. A rater's goals or intentions in providing performance ratings are usually much more complex than the desire to provide an accurate rating on performance behavior (Yammarino & Atwater, 2001). These findings suggest that rather than providing more reliable ratings of leader behavior, aggregated competency scores are, in fact, an aggregate of personal motivations and artifacts of the leader-rater relationship. Although it might make intuitive sense to aggregate scores within rater groups to provide a more objective rating of leader behavior as seen from a particular organizational perspective, it may also be the case that disagreement observed among rater groups is simply a byproduct of the disagreement that exists among raters in general. Therefore, an analysis that can separate out some of the social dynamics within the rating group may be more beneficial for leader development.

The present study does not argue against the practice of aggregating multiple ratings as part of leader evaluation. Rather, the goal is to explore the potential of using a complementary technique for analyzing MSF data that can provide additional insights into leadership, and which may be as relevant to leader development as those of aggregated competency scores are to leader evaluation.

Relational analysis for MSF. Gooty and Yammarino (2011) note that studies focusing on interpersonal dyads, including relationships between leaders and followers, have received

much less attention than the study of individuals or larger groups. A lack of methods for analyzing dyadic data at this scale may partially explain why MSF scoring methods remain focused on conventional rating aggregation; however, more sophisticated techniques are now widely available and are becoming increasingly simple to implement. The Social Relations Model (SRM; Malloy & Kenny, 1986; Kenny, 1994) provides a theory and method of measuring interpersonal interaction and group dynamics that may offer valuable insights when applied to MSF ratings in order to predict leader effectiveness outcomes.

An early introduction of the SRM into organizational research applied this versatile model to peer evaluation data from student teams (Greguras et al., 2001). The SRM analyzes nonindependence among individuals within the same group and considers the dyadic relationships among them. The SRM can be considered a special case of G Theory (Cronbach et al., 1972). However, analyses of the SRM extend beyond G Theory, because once variance components are isolated, they are correlated with each other and with leader self-ratings to provide insights into dyadic relationships within the group and the overall social dynamics.

The general assumption of the SRM is that individuals who interact on a regular basis are the best judges of each other's behaviors (Kenny, 1994). Although G Theory has previously been applied to multi-source peer ratings similar to MSF assessments (Greguras & Robie, 1998), to my knowledge components of the SRM calculated from MSF ratings have not previously been considered. Moreover, these studies that utilize G Theory do not use variance components to predict manager effectiveness outcomes. The SRM measures three types of variance among ratings: *target* (i.e., dispositional/ consensus) variance, *perceiver* (i.e., situational/assimilation) variance, and *relationship* (i.e., interactive/unique relations) variance. Each of these components on their own can provide valuable information regarding manager interactions with raters. *Target*

variance is a measure of consensus, or how consistently others provide similar ratings on a single leader, which could be an indication of how strong the leader's standing on a competency is. *Perceiver* variance is a measure of assimilation, or how alike a single leader perceives others to be. Finally, *relationship* variance measures the interaction between target and perceiver variance, which indicates whether ratings are primarily a function of the specific dyadic influence between two raters. Analyzing the combination of all three variance components provides potentially powerful insights to understanding social dynamics that underlie leader behaviors in the context of their rating group. Knowing how leaders rate each other can potentially provide leaders with valuable feedback regarding their own self-regulation practices.

Relationship variance provides a measure of unique, reciprocal relationships among raters, which could be an indication of the high quality, dyadic relationships between leaders and others mentioned previously. Therefore, response patterns between leaders and others could provide circumstantial evidence of underlying social processes and self-regulation within the group. The way in which raters respond to questions about the leader is based on each rater's individual experiences with the leader, which is likely to be directly related to critical aspects of the leader competency model and overall effectiveness. Research has already demonstrated that components of the SRM (i.e., *target*, *perceiver*, *relationship* variance) are related to team processes and team effectiveness outcomes (LeDoux, Gorman, & Woehr, 2012). Therefore, it is likely that these same components measured within rater groups on MSF may provide key information regarding relational processes within the group that are likely to be strongly tied to overall leader effectiveness.

Hypotheses. The prior research outlined above provides compelling evidence that variance in ratings will be directly related to various measures of leader effectiveness including

performance, promotions, and manager quality. However, the interactionist perspective of leadership suggests that some rating variance may be differentially predictive of various leader competencies. The MSF assessment used in the present study includes items measuring leader competencies that are related to social relationships, including their ability to collaborate and help subordinates acquire new skills in addition to having business-oriented abilities such as supporting larger organizational objectives. The arguments made in this proposal thus far recommend that the analysis used to measure MSF data should better align with the leadership theory underlying the purported use of the MSF ratings. At the time of data collection, the target organization rated leaders separately on business and interpersonal-related objectives as part of their regular performance appraisal process, which provides an interesting opportunity for analysis of the validity of the SRM components relative to leader effectiveness through examination of the relational processes inherent to the rating group. This same logic suggests that aggregated scores may be the best predictors of business-related performance while dyadic variance components may be the best predictors of interpersonal-related performance. The following hypotheses focus on the *middle tier* leader assessments since this is the largest population of leaders that will yield the highest number of round-robin groups.

SRM-H1: (a) Target, (b) Partner, and (c) Relationship variance components each account for a significant portion of total variance in peer ratings of leader behavior on the middle tier *assessments.*

SRM-H2: Target variance will be highest for the task-related competency compared to (a) partner and (b) relationship variance on middle tier *assessments.*

SRM-H3: Relationship variance will be highest for the socio-emotional competency compared to (a) target and (b) partner variance on middle tier *assessments.*

Part III: Rating Agreement and Leader Effectiveness

Discrepancy between self and other ratings is another conventional application of MSF that is generally considered an indicator of leader self-awareness, which has been linked to various leader effectiveness outcomes (Church, 1994; Fleenor, Smither, Atwater, Braddy, & Sturm, 2010). Grounded in Self-Insight Theory (Allport, 1937), MSF leader development programs adhere to the notion that leaders are motivated to change behaviors in areas where they see discrepancy between self- and other-ratings. This focus on discrepancy has led to MSF assessment research investigating potential links between self-other agreement (SOA) and leader effectiveness outcomes. In their generally accepted framework, Atwater and Yammarino (1997) identify leaders as fitting into one of four distinct categories of SOA: over-estimation, under-estimation, agree-high, and agree-low. Atwater and Yammarino (1997) proposed a relative categorization scheme that first created a distribution of all dyadic discrepancy scores and then categorized leaders with an average discrepancy over half a standard deviation from the mean discrepancy as over-estimators. Under-estimators were identified as having an average group discrepancy half a standard deviation below the mean. Leaders in the middle were considered either agree-high (i.e., above the mean) or agree-low leaders (i.e., below the mean). This categorization technique addressed two important issues regarding leader self-ratings that were prevalent in the literature at the time: (1) problems with using absolute discrepancy scores, and (2) inflation bias of self-ratings.

In their seminal paper on SOA, Atwater and Yammarino (1992) cite evidence from studies finding that leniency bias (i.e., under-estimation) in leader self-ratings was positively related to self-esteem (Farh & Dobbins, 1989; Goffin & Anderson, 2007) and demonstrated a positive relationship with private self-consciousness (i.e., self-awareness; Froming & Carver,

1981). Furthermore, researchers employing this categorization system have generally corroborated this with findings indicating that over-estimators, or leaders who rate themselves higher than how others rate them, have lower levels of performance compared to leaders with self-ratings that agreement with others' ratings (Bass & Yammarino, 1991; Carless, Mann, & Wearing, 1998) and that leaders who under-rate their own performance may be more effective (Amundsen & Martinsen, 2014). In other words, MSF researchers generally agree that leaders who inflate, or enhance, self-ratings of performance tend to be less effective leaders compared to those who agree with others or under-rate their ability.

A significant concern with conventional SOA research in MSF is that it does not fully consider implications of the basic human tendency to bias ratings of the self (Greenwald, 1980; Johnston, 1967; Taylor & Brown, 1988). Even though Atwater and Yammarino (1997) sought to reduce the influence of leader self-rating bias by using a distribution of discrepancy scores to determine categories, SOA research in general almost exclusively considers absolute self-ratings as they relate to others' absolute ratings of the leaders. This conventional approach potentially clouds the relationship between discrepancy and leader effectiveness outcomes because it fails to consider two important aspects of leader self-ratings: (1) that, generally, individuals possess a higher esteem for the self than others do, and (2) how leaders view their own ability compared to the ability of other leaders on the same MSF assessments. In other words, the one-sided SOA categories that only compare leader and other ratings may only provide limited information for leader development, and including the two-way ratings derived from the SRM may augment MSF in a way that makes it more relevant for leadership development.

Despite the attractive simplicity of Atwater and Yammarino's (1997) categorization model, some researchers were skeptical about the validity of these SOA categories for predicting

leader effectiveness outcomes. To supplement the four-category model of SOA, Fleenor, McCauley and Brutus (1996) compared these categories to a proposed six-category model that further differentiated the over- and under-estimator categories originally described by Atwater & Yammarino (1997) into over-/under-estimators with high average scores and over-/under-estimators with low average scores. Using the four-category (Atwater & Yammarino, 1997) model, the agree-high and the under-estimator categories were the best predictors of leader effectiveness; however, the six-category (Fleenor et al., 1996) model indicated that agree high and the over-estimator high group were the better predictors. In both categorization schemes, leaders with high average scores and no discrepancy between ratings were generally considered more effective; however, results from the six-category model call into question the validity of the claim that leaders who under-estimate are more effective than leaders who over-estimate. Further supporting this controversy, Brutus, Fleenor, and Tisak (1999) found that variability in self-ratings did not explain significant variance in leader effectiveness outcomes, rather, peer ratings proved the stronger predictor. These findings are further supported by more recent research that found no relationship between self-ratings and follower-reported LMX relationship quality (Barbuto, Wilmot, & Story, 2011). In light of these findings, Barbuto et al. (2011) concluded that categorization methods do not fully explain variability in leader outcomes.

Part III of the present study digs deeper into the potential for MSF assessment ratings in order to better understand leader outcomes based on the premise of leader self-regulation. By considering other theories of self-perception as well as score discrepancies in MSF assessments between leader self-ratings and ratings of others, application of the Social Relations Model may provide more useful metrics regarding leader self-awareness, and that may be better predictors of leader effectiveness outcomes.

Theories of self-perception. Leader self-ratings have historically been incorporated into MSF assessment processes as well as leader development programs as indicators of self-awareness, which is measured in terms of rating agreement. In fact, the goal of many MSF development programs is to reduce the gap between self- and other-ratings. Years of research have shown that self-awareness is important for leader development; however, this relatively limited view of what the MSF ratings mean in the context of leader self-awareness ignores other factors that may also provide valuable feedback when the goal is leadership development. While Atwater and Yammarino (1992) acknowledge the existence of alternative theories regarding individual characteristics and their influence on self-ratings, in fact an entire body of psychological literature is dedicated to individuals' general tendency to inflate one's perception of the self (Greenwald, 1980; Johnston, 1967; Taylor & Brown, 1988). These studies generally find that the tendency to inflate self-ratings, henceforth referred to in the current proposal as the self-serving bias (SSB), is positively related to self-esteem and general wellbeing.

Self-Insight Theory (Allport, 1937) supports the idea that agreement between self and other ratings could provide an indication of leader self-awareness; however, this is only one aspect of self-perception. In the context of MSF, agreement between self- and other-ratings may just as likely be an indication that the leader demonstrates humility and camaraderie with followers. In contrast, Social Comparison Theory (Festinger, 1954) considers how the leader rates him/herself compared to how he/she rates others. In the context of MSF, the social comparison model invites the possibility that higher ratings may suggest that a leader believes he/she is capable of performing at a higher level as compared to other leaders, which could be an indication of leader self-confidence (Hollenbeck & Hall, 2004). Some research in self-perception has taken Social Comparison Theory into account (Kwan, John, Kenny, Bond, Robins, 2004) but

a recent review of SOA research acknowledges that MSF research primarily relies on Self-Insight Theory, ignoring the potential impact of social comparison.

As stated in the previous section, a combination of observed leader behaviors and the individual motivations of the raters are known to influence MSF ratings. In their recent review of self-perception research, Fleenor et al. (2010) acknowledge that this is also true for leaders when they are rating themselves and rating others. Therefore, variability in MSF self-ratings is just as likely to represent individual characteristics of the leader, such as self-esteem or emotional intelligence, as they are to represent full awareness and acknowledgement of the self. In other words, more savvy leaders may have an acute sense of awareness of how others perceive them, even though they may not believe that this is an accurate representation of their behavior (Pronin, Gilovich, & Ross, 2004). Furthermore, self-perception research also finds that how individuals rate themselves has an impact on how they rate others (Shore, Adams, & Tashchian, 1998). Therefore, combining leader self-ratings with the round-robin ratings in the SRM, which considers both the social-comparison and self-insight models simultaneously, may provide leaders with a more complete picture of their self-regulation tendencies based on self-ratings, and may also be a more appropriate metric for use in leadership development programs.

Leader Self-Enhancement

Self-enhancement is a broad term in the literature that refers to the human tendency to hold a subjectively favorable view of one's self (Sedikides & Gregg, 2008). In their relatively recent review of the topic, Sedikides and Gregg (2008) discuss the various streams of research that consider the possible consequences of individuals holding an inflated sense of self and the increased probability of idiosyncratic findings that could result from researchers only considering self-perception from a self-insights perspective. In fact, their review specifically

mentions the ongoing debate regarding the theories of self-perception mentioned in the previous section, and they acknowledge that only considering leader self-ratings in comparison to how others rate the leader is a very limited view of self-enhancement. Therefore, it should not be surprising that leader self-awareness, as conventionally measured with MSF assessment data, is not a stable predictor of leader effectiveness. Incorporating leader self-ratings into a relational model such as the SRM is an improvement over conventional SOA because it allows for the simultaneous measurement of self-insight and social comparison in regards to leader ratings (Kenny & West, 2010; Kwan, et al., 2004). This final section of the present study combines insights from previous work that considers self-ratings from an SRM perspective to measure self-enhancement in a unique way. This technique may provide a more comprehensive understanding of leader self-regulation associated with patterns of leader self-ratings; specifically, whether leader inflation of self-ratings could be an indicator of self-confidence in some cases.

Yammarino (2003) emphasizes the importance of dyadic relationships in the study of leadership and that MSF ratings may provide important information about the individuals who are providing the ratings as opposed to the behaviors of the focal leaders themselves. This logic suggests that conventional SOA measures might provide information more relevant to those providing the ratings rather than information specific to the leader being rated. Therefore, participating leaders' ratings of others who are also rating them could be harnessed to provide some additional information regarding that leader. For example, a leader identified as an under-estimator using conventional MSF agreement measures may provide ratings for others that are lower than those he or she provides for him or herself. In this case, the leader may be more appropriately identified as an over-estimator because he or she systematically rates others lower

than his or her self-ratings. Only considering the one-way dyadic ratings in conventional SOA may miss potentially valuable information regarding leader behavior and leader self-regulation because internal motivations are likely to be different for leaders depending on whether they see others as generally more or less favorable than they seem themselves.

In order to diagnose deficiencies in leader behavior, we cannot simply look at discrepancy because this does not take into account potential response biases on the part of a self- versus other-rating. A leader rating his or herself will give a relatively positive score in line with his/her self-esteem but that same leader may also recognize the superiority of his/her peers compared to his/her own abilities. So if the leader rates him/herself higher than others' ratings, but the self-rating is still lower than how the leader rates others, a traditional SOA category approach would exaggerate the leader's overall rating tendencies because this leader would be considered an over-estimator even though he/she rates others higher than his/her self.

In their study on self-enhancement, Kwan and her colleagues (2004) compared methods of calculating a single measure of self-enhancement as a combination of a self-insight index and a social comparison index. In their proposed model, the two indices are defined as *k* and *q*. The *k* parameter provides the social comparison index and measures the effect of the *perceiver* effect (i.e., how leaders rate others) on leader self-ratings. The *q* parameter provides the self-insight index and measures the effect of the *target* effect (i.e., how others rate the leader) on leader self-ratings. Conventional SOA analyses for MSF assessment data focus exclusively on the *q* parameter; therefore, also including the *k* parameter when calculating a measure of self-enhancement provides a novel assessment of leader self-awareness. In support of this two-factor model, Kwan and colleagues (2004) found that including the *k* parameter in a measure of self-

enhancement was a stronger predictor of individual task performance and individual self-esteem than when *q* or *k* were tested as separate predictors.

The predictive relationships reported by Kwan et al. (2004) provide evidence in support of including the *k* parameter along with the *q* parameter in measures of leader self-enhancement. In other words, a leader who, from a self-insight perspective, over estimates his or her self-ratings but also rates others at the same level or higher, may not be self-enhancing to the same extent as an over-estimator who always rates others lower than himself. However, one drawback to the method used by Kwan et al. (2004) is that it relies on absolute discrepancies calculated by simply subtracting the *perceiver* and *target* effects from leader self-ratings. Calculating absolute discrepancies in this way are known to be problematic because *k* and *q* are assumed to have an equal impact on leader self-ratings. Kenny and West (2010) recently challenged this assumption, and in a meta-analysis of 24 different studies that included both self and other ratings demonstrated that *k* and *q* are far more variable than has previously been assumed in self-ratings research. Therefore, a comprehensive model of self-enhancement would ideally account for the influence of these two parameters on self-ratings without constraining them to be equal. To the author's knowledge, the present study is the first to estimate these parameters separately in a full model of leader self-enhancement, and to do so with data collected from an actual work organization.

Hypotheses. The arguments posed above assume that self-enhancement measured through the SRM is less prone to bias found in more conventional SOA discrepancy measures. Therefore, self-enhancement as measured in the present study is expected to be a better predictor of leader effectiveness outcomes compared to the over-rater and under-rater categories

conventionally used with MSF. The following hypotheses focus on *middle tier* leaders since this population is the largest and most stable over time.

SE-H1: Self-enhancement of middle tier *leaders assessed via SRM will be negatively related to leader quality.*

SE-H2: Self-enhancement of middle tier *assessed via SRM will be positively related to leader termination.*

SE-H3: Self-enhancement of middle tier *assessed via SRM will be negatively related to leader promotions.*

SE-H4: Self-enhancement of middle tier *assessed via SRM will be negatively related to leader performance.*

Method

Participants

As mentioned previously, the present field study uses archival data from a large consumer goods organization collected between 2009-2014. Part I of the dissertation will include all leaders who completed any of the MSF assessments (i.e., *top tier* leaders, *middle tier* leaders, and *foundational tier* leaders) during this time period. Since 2009, 725 *top tier* leaders (72% male), 5,753 *middle tier* leaders (69% male), and 2,819 *foundational tier* leaders (67% male) completed MSF assessments. Overall, leaders at the time of the assessment had been working at the organization for approximately 14 years (M= 13.60; SD= 7.36, Min= 2.59, Max=44.48). A majority of leaders are based in the US (57%) and the sample includes leaders from each global region (59% North/Central/South America, 24% Europe, 17% Middle East/Asia/Africa) and each major business function (7.2% IT, 10.9% Marketing, 15.1% Finance, 7.6% Human Resources, 19.9% Operations, 8.2% R&D, 19.6% Sales, 11.5% Other). Including all records of

the assessment since 2009 means that some participants will have multiple records included in IRT analyses. Crossover in assessments may result in some demographic distortion; demographic information is reflective of when the participant completed the assessment.

Parts II and III of the study include a subset of the larger population who completed MSF assessments in 2013 and 2014. Due to potential confounds regarding rater source, the round-robin dataset only considered ratings made by leader peers. In this way, all members of the round-robin groups are leaders with similar levels of responsibility in the present organization. Criteria for round robin groups included in the final dataset are as follows: (1) at least two members of each group provided valid self-ratings, (2) the full round-robin group consisted of at least 3 individuals, and (3) all ratings were made in the same calendar year. A custom FORTRAN program was used to apply the above criteria to the present dataset, which resulted in the loss of approximately 12% of ratings from the full 2013-2014 dataset. Additionally, this final dataset no longer resembles the one-way dyadic ratings from the original dataset because all raters who meet the criteria are represented as focal leaders in the original data. Therefore, all participants in Part II and III are leaders who provided self and peer ratings in the organization's MSF assessment as a *middle tier* leader within each year from 2013-2014.

The final dataset for Parts II and III includes 1,078 peer leaders included in 351 MSF round-robin groups; the majority of groups contain 3 peer leaders (N = 328), a smaller number of groups contain 4 peer leaders (N = 21), and very few contain 5 peer leaders (N = 2). Only 5% of groups (N = 18) have complete round-robin data with an average of 32% missing data across the remaining groups. Target ratings are missing for 61 leaders across 50 groups and perceiver ratings are missing for 285 leaders across 101 groups. The demographic information for the leaders included in this dataset mirrors that of the broader organization and those reported above

for Part I. Furthermore, MSF assessment scores for the leaders included in the round robin dataset are not statistically different from organization norms. Norms are calculated within each assessment group (i.e., leaders, senior leaders, and all employees) using a 3-year rolling average of scores for each competency. The most recent norms for the *middle tier* leader group used in Parts II and III of the dissertation are 3.87 for task-related skills and 3.90 for socio-emotional skills on a 5-point scale. In the final round-robin dataset, leader norms are 3.82 for task-related skills and 3.85. Therefore, the leaders in that dataset included in Parts II and III of the present study are representative of the broader leader population who participate in MSF assessments.

Procedure

The organization regularly collects MSF assessment data to accommodate the needs of the various business units and functions at multiple intervals throughout each year. A majority of the data represented in the present study was collected during the regular annual administration held in the 3rd Quarter of the fiscal year. However, smaller data collection cycles were also implemented throughout the year to accommodate less common requests and to collect assessment data for any employees participating in business- or function-specific development programs. These programs are in addition to the mainstream corporate development process, and even though smaller business units administer tailored development programs to their employees, the organization's general framework for assessing and providing feedback is managed through the corporate office and remains relatively standardized worldwide.

In addition to standardization in assessing and providing feedback to employees, the administration with which employees participate in MSF assessments is also standardized throughout the corporation. The criteria for employee eligibility for participating in a MSF assessment are: (1) at least 18 months have passed since a previous assessment, (2) the employee

has spent at least 6 months in a new role, and/or (3) the employee has been invited to participate in a tailored leadership program. The above conditions apply to all leaders participating in MSF assessments; therefore, all available MSF data were considered for analysis.

Measures

An experienced internal assessment team developed the MSF assessment used in the present study. Therefore, the leadership dimensions and competencies measured by the assessment directly map to the overall leadership and career model used within the organization. Aside from the MSF items and test scores, various outcome measures described below are also central to the hypotheses being tested in the present study to understand how MSF assessment ratings influence leader effectiveness.

Multi-source feedback assessment. The MSF assessment used in the present study was custom-built for the present organization based on the leadership competency model. The dimensions used in the present study consist of a core assessment, which includes 23 items that are the same for every leader participating in the assessment, regardless of the leader's level in the organization. The MSF assessments used for junior and senior levels each include an additional 24 items that are specific to the leader's level in the organization. Regardless of assessment type, all items considered in the present study measure various aspects of leader behavior including: strategy execution, communication skills, and interpersonal development. Items ask raters to indicate the extent to which each leader participates in relevant behaviors. Sample behavior items include: "Demonstrates perseverance and resilience in the pursuit of goals" and "Treats all people with respect and fairness". Exploratory and confirmatory factor analyses were conducted to determine the most appropriate factor structure for the items. These factors were used in subsequent analyses.

Leader performance. Leader performance is rated on an annual basis at the present organization. Leaders receive two ratings of performance: one rating on task-related objectives and one rating on socio-emotional objectives. Ratings are each made on a 5-point rating scale, are largely determined by each leader's supervisor and are based on the leader's successful execution of personal objectives that were set in the beginning of the year. Once assigned, each of the two ratings is then calibrated for the entire organization to ensure a normal distribution. Performance on business outcomes and interpersonal outcomes will be used separately in the present study since these were designed to capture two distinct domains. The two ratings are also relevant to the distinction being made regarding individual differences in leader ability and traits (i.e., task-related performance) as well as relational processes of leadership (i.e., socio-emotional performance).

Leader quality. Leaders with three or more direct subordinates at the present organization are rated on an annual basis on the overall quality of their leadership. This 12-item assessment is separate from the interpersonal dimensions of the MSF assessments and captures a unidimensional measure of overall leader quality. Items ask subordinates to comment specifically on how trustworthy the leader is, on the leader's communication skills, and on the leader's ability to provide effective feedback, along with other important aspects related to the overall quality of the leader. Sample items include: "Gives others full credit for their ideas and contributions" and "Constructively addresses performance issues". Similar to the MSF assessment, the manager quality assessment was also custom-built for use in the present organization.

Termination. The organization's turnover data were used to document which leaders in the dataset were involuntarily terminated at the present organization. Data in the present study

were collected over a six-year period, and this made it likely that some leaders were no longer employed at the company at the end of this period given the average turnover rate of 4% annually. Additionally, leaders with high levels of self-enhancement in this organization were generally judged to not take development activities seriously, and this may have negatively impacted their standing in the organization. Involuntary termination data is therefore a useful metric for the present study because these data provide some indication that leaders' leadership styles did not fit the needs of the organization, or that the organization was unwilling to support the needs of these leaders.

Promotions. The present study also uses the promotion data from the organization. Whether MSF ratings are predictive of leader success in the organization can be partly determined by the extent to which leaders with high MSF ratings are promoted within the organization. Therefore, the present study utilizes promotion data to determine which leaders are successfully advancing their careers. As a contrast to turnover data, this metric supplies a relatively objective measure of how effective each leader is since leaders who are more effective overall should presumably experience a higher promotion rate than those who are less effective.

Tenure. Leader tenure was included as a control variable in SRM analyses. Researchers acknowledge that high-quality, reciprocal dyadic relationships with leaders take time to fully develop (Graen, Hui, & Taylor, 2006) and longitudinal studies have shown that tenure is a strong predictor of relationship quality over time (Sin, Nahrgang, & Moregeson, 2009). The amount of time the leader has spent cultivating relationships with employees is relevant for both types of analyses used in the present study. First, the length of time a leader has spent in the organization will impact how much influence the leadership competency model or internal leadership training and messaging have on the leader's perception of the behaviors being rated on the MSF

assessment. Tenure is also important for the SRM since this model explicitly measures dyadic influences within a group and a key assumption of the model is that close others have the best knowledge of a leader's capability (Kenny, 1994). Furthermore, researchers have found that the length of job experience has an effect on individual ratings of others (Cascio & Valenzi, 1977). Leaders with shorter tenure may not have had enough time to demonstrate their full capability to raters, which could bias MSF ratings.

Gender. Leader gender was also included in SRM analyses as a control variable. Gender is commonly known to be of concern in leadership research due to gender stereotyping that influences how male leaders are perceived compared to female leaders. Gender is also relevant to the present study because men demonstrate more bias in self-ratings than females (Fleenor et al., 2010). Therefore, gender is important to include in analyses as it may explain some variance in ratings.

Results

Power Analyses

The data available in the present study provide an ample number of data points for hypothesis testing. Part I of the present study utilizes IRT to investigate parameter drift among MSF assessment items. Researchers have shown that IRT parameters for a polytomous model, such as the GRM being used in the present study, can be estimated with a sample that includes 2,400 – 3,000 or more individuals (Drasgow, Chernyshenko, & Stark, 2010; Hambleton, Swaminathan, & Rogers, 1991). The archival data from the present organization exceeds this minimum requirement for all three of the tiered MSF assessments. The *top tier* pool is the smallest with 725 leaders; however, with an organizational minimum requirement of 3 raters per leader, there are at least 2,175 available ratings. The *middle tier* pool is the largest with 5,753

leaders and a minimum of 17,259 ratings. Therefore, the archival sample provides ample power for estimating all parameters of the GRM and detecting parameter drift.

Parts II and III of the present study require the estimation of several variance components as part of the SRM. In their comprehensive guide on the subject, Kenny and colleagues (2006, p. 215) report that a minimum of 17, four-person round-robin groups are necessary to confidently estimate all parameters of the SRM. As mentioned previously, the dataset for Parts II and III of the present study include 351 round-robin groups and 23 of those groups contain 4 or more members, which meets the minimum requirement. Despite losing a large amount of MSF assessment data during the creation of the round-robin groups, the remaining sample still provides an adequate sample size to confidently score all SRM parameters.

Preliminary Analyses

Part I. Exploratory factor analysis (EFA) was conducted separately for items included on each dimension of the MSF assessment in order to establish unidimensionality among the task-related items and the socio-emotional items. This process was purposefully more liberal than the preliminary analyses conducted for Parts II and III of the study and only meant to confirm that data meet the assumption of IRT analyses that requires the use of unidimensional factors. Given that the MSF assessment purposefully splits task-related and socio-emotional items into two separate dimension scores, it was assumed that each item set would load onto one primary factor within each leadership tier. Previous work suggests that single factors explaining at least 20% of variance can be considered unidimensional for the purposes of IRT (Reckase, 1979). Therefore, EFA analyses were conducted with each of the tiered leadership assessments in order to establish unidimensionality for task-related and socio-emotional factors, which resulted in six single-

factor solutions for task-related and socio-emotional item sets for *foundational*, *middle*, and *top tier* assessments.

As mentioned previously, the *middle tier* assessment was used to set the baseline for the preliminary analyses, and is the sole focus of analyses conducted in Parts II and III of the present study. Therefore, items from this assessment were used to establish the initial factor structure, which was then replicated for items from the *foundational* and *top tier* assessments. Results confirmed that a single-factor solution was appropriate for both dimensions on the *middle tier* assessment with the task-related and socio-emotional factors each explaining over 50% of variance (Table 1). Similar factor structures emerged for the *foundational* and *top tier* assessments with both dimensions explaining at least 60% and 51% of variance, respectively (Table 1).

Parts II and III. Parts II and III of the present study require dimension scores to serve as outcome variables for the SRM and self-enhancement analyses. As mentioned previously, these analyses focus on a subset of the *middle tier* MSF assessment data using only peer ratings. Therefore, factor analyses were conducted with peer ratings for *middle tier* leaders. Factor structure was investigated using the two halves of the larger *middle tier* data set including all peer ratings from 2009-2014.

Factor structure was determined by using parallel analysis followed by an EFA using principle axis factoring with an oblique rotated solution. The number of factors to retain was determined based on scree plots from the parallel analysis as well as variance accounted for by the EFA. The scree plots (Figures 1 and 2) indicated that the MSF items split into approximately 7 factors; however, a majority of the items loaded onto the first two factors (Table 2). Additionally, previous work on psychometric analyses of MSF assessment data supports a two-

factor solution regardless of the number of smaller factors typically found in EFA of MSF assessment data (Craig & Kaiser, 2003). The present study adopted this strategy and retained items with the strongest loadings across the task-related and socio-emotional items. The first factor explained 67% of variance in responses and included all task-related items as well as six items related to leader development from the original socio-emotional dimension. The second factor explained 33% of variance and included 12 items from the original socio-emotional assessment. The top 14 items loading on the first factor were retained to have an equal number of items for each dimension.

To empirically test how well the 2-factor solution identified by the parallel analysis and EFA fits the data, confirmatory factor analysis (CFA) was applied to the second random half data set. Fit statistics (Table 3) confirmed the 2-factor solution for peer ratings. Finally, reliability analyses also support each dimension of the 2-factor structure identified by the CFA model. The items demonstrated very good internal reliability for peer ratings of task-related (14 items; a= .93) and socio-emotional (12 items; a= .94) leadership dimensions. Ratings from items retained in the two main factors were averaged to provide final task-related and socio-emotional scores, which were used as outcomes for analyses in Parts II and III of the present study.

Part I – IRT Analyses

The goal of Part I was to utilize IRT methods to understand whether MSF assessment item parameters and test information change over time within the same leader population. The procedures described below are based on recommendations for conducting IRT in a MSF assessment framework. Craig and Kaiser (2001; 2003) suggest using all respondents within a single rating category to avoid potential confounds due to rater type; therefore, analyses in the present study will be conducted separately within each rating group (i.e., manager, peer, and

direct report ratings). Self-ratings were not used in the analyses for Part I because ratings from others are what the present organization primarily uses to differentiate leaders on the various dimensions of leadership. In contrast, self-ratings are used as a comparison tool to identify areas where leader self-awareness may be lacking. Leader self-ratings will be examined more fully in Part III of this study.

Model fit. As stated above, preliminary analyses support a unidimensional factor solution for the items included on the task-related and socio-emotional MSF assessment dimensions, respectively. Therefore, all items were included in initial tests of model fit. Items for each dimension were fitted to the graded response model (GRM; Samajima, 1969), which is an IRT model that can handle polytomous response options such as the 5-point Likert scale used on the MSF assessment. The GRM fit the data reasonably well for *middle tier* task-related ($RMSEA_{Peer}$= .07; $RMSEA_{DR}$= .06; $RMSEA_{Mgr}$= .07) as well as socio-emotional items ($RMSEA_{Peer}$= .07; $RMSEA_{DR}$= .07; $RMSEA_{Mgr}$= .08) and fit the data reasonably well for *top tier* task-related ($RMSEA_{Peer}$= .05; $RMSEA_{DR}$= .07; $RMSEA_{Mgr}$= .08) as well as socio-emotional items ($RMSEA_{Peer}$= .07; $RMSEA_{DR}$= .06; $RMSEA_{Mgr}$= .07). Unfortunately, the GRM did not fit the data appropriately for *foundational tier* leaders (a full list of fit statistics are available in Table 4). Poor model fit for *foundational tier* leaders is likely due to the smaller sample size and fewer items available for this MSF assessment, which prevented the GRM from converging in some cases and resulted in extremely poor fit in others (Table 4). As a result, items from the *foundational tier* MSF assessment were excluded from future analyses.

It may be worth stating here that the goal of using IRT in Part I of the present study was to empirically address theoretical questions regarding the assessment and not for scale development. Therefore, analyses can proceed without any concern for moderate to high model

fit in the case of some rating sources. Ratings for *foundational tier* leaders were excluded because these models either did not converge or fit so poorly that accurate interpretation of the results would not have been possible.

Item discrimination. As an initial step in the psychometric evaluation of the MSF assessment items in the present study, item discrimination parameters were assessed to determine whether items written specifically for the tiered leadership competencies of task-related and socio-emotional leadership behaviors actually provide better discrimination among leaders as compared to core items, which are written more broadly about foundational leadership behaviors. The first set of hypotheses tests whether items written specifically for each leadership tier (i.e., *foundational, middle,* and *top tier*) provide better discrimination compared to core items. To avoid potential confounds from repeat assessments, all ratings included in analyses of *IRT-H1* and *IRT-H2* only include the first assessment that each focal leader participated in within each leadership tier.

Hypothesis *IRT-H1* tested whether core items had higher discrimination for *foundational tier* leaders compared to *middle* and *top tier* leaders. As mentioned previously, ratings provided for *foundational tier* leaders did not fit the hypothesized model for the data, so item parameters such as item discrimination cannot be interpreted. Therefore, *IRT-H1* was not supported.

Hypothesis *IRT-H2* tested whether the respective tailored items provide higher discrimination for *middle* and *top tier* leaders. To test *IRT-H2,* the item discrimination parameters generated by the GRM were ranked from highest to lowest (Tables 5 and 6). Once items were sorted, the percentage of tailored items in the first half of the item set were calculated in order to determine whether tailored items provide relatively higher discrimination of leader behavior compared to core items. Results at the top of Table 5 indicate that for *middle tier*

assessment task-related items, the tailored items provide relatively more discrimination for peer (70%) and manager ratings (60%), but for direct report ratings there were an even number of core and tailored items providing the highest discrimination. For socio-emotional items, tailored items provided slightly higher discrimination for all rating sources (54%). Therefore, *IRT-H2a* is partially supported.

Tests of *IRT-H2b* indicate that for *top tier* leaders tailored items provide relatively higher discrimination across all rating sources for task-related ratings. However, tailored items provide much lower discrimination across all rating sources for socio-emotional items (Table 6). Thus, *IRT-H2b* is partially supported.

Changes over Time

The second goal of Part I of the present study was to evaluate how MSF assessment data changes over time. In order to interpret test information curves across years, the same leaders must be included in each dataset. Only including leaders who repeated the assessment at the same interval allows for the interpretation of assessment parameters across multiple years in order to eliminate any potential confounds due to changes in the leader sample. The hypotheses originally were written to test parameter change over a period of five years; however, a closer analysis of repeated leaders within the overall dataset indicated that a majority of leaders repeated the MSF assessment in 2013 after initially completing the assessment in 2009. These were the only two years with a large enough sample size of identical leaders with ratings that could be analyzed and interpreted across years and also within rating groups (i.e., peer and direct report). Manager ratings were not used in these analyses because the sample was too small for IRT models to converge. The final dataset includes Time 1 (2009) and Time 2 (2013) data for

210 leaders with an average of 4 peer ratings and 5 direct report ratings per leader. These datasets were used to test the remaining hypotheses in Part I.

Test information. Hypothesis *IRT-H3* tests whether continued use of MSF assessments results in a reduction of the amount of information provided by the assessment. The GRM was fit to ratings provided by peers and direct reports in 2009 and then again in 2013. The test information curve was then plotted for each model. Results for the task-related dimension are depicted for peers in Figure 3 and for direct reports in Figure 4. Results for the socio-emotional dimension are depicted for peers in Figure 5 and for direct reports in Figure 6. Hypothesis *IRT-H3* predicts that test information will degrade over time such that information is higher in 2009 than it is for 2013. Results partially support *IRT-H3* for the task-related dimension with peer ratings providing less information in 2013 than they did in 2009. However, direct report ratings have the opposite pattern providing more information in 2013 than in 2009, which may be some indication of the amount of time the raters have known the leader. Therefore, *IRT-H3* is partially supported for the task-related dimension.

Results for *IRT-H3* demonstrate a much more complicated shift in ratings over time for the socio-emotional dimension. Overall, the results show that for the socio-emotional dimension of leadership, *IRT-H3* was not supported. Peer ratings seem to provide similar information from 2009 to 2013; however, the location of this information appears to shift with peers providing more information for higher ability leaders in 2009 compared to 2013. Similar to the task-related results, direct report ratings provided more information regarding leader behavior in 2013 as compared to 2009. Again, results may be indicative of historical shifts in the relationship between the leader and raters over a period of four years.

Differential item functioning. The final stage of Part I involves testing whether parameters of MSF assessment items remain stable over time within rating source. The present organization uses individual item ratings to identify areas where a leader may need to participate in more development activities; therefore, whether these items function the same over time is important to consider for leaders who are rated multiple times over several years. Changes in item parameters can be either uniform or non-uniform, which could result in problematic interpretation of a leader's progress over time. For example, uniform DIF occurs when there are systematic changes in the likelihood of a rater selecting a response option across the entire range of leader ability. In other words, uniform DIF would occur if response categories in 2013 refer to consistently higher leader ability compared to 2009. This type of DIF could cause a leader to assume that no progress has been made in certain development areas when in fact, progress has been made but raters are systematically using the response options in a different way than in previous years. Non-uniform DIF can happen in one of two ways: (1) non-crossing non-uniform DIF is when the likelihood of choosing a response option changes across all options, but at different degrees, and (2) crossing non-uniform DIF is when the likelihood of response options changes for some, but not all response options.

Hypothesis *IRT-H4* tests for DIF over time using the calculation methods outlined in the *mirt* package available in R (Chalmers, 2012). As before, IRT analyses used the graded response model and assumed unidimensionality for both the task-related and socio-emotional MSF assessment dimensions, which satisfy the first two steps in calculating DIF (Collins, Raju, & Edwards, 2000). The third step necessary to appropriately calculate DIF involves the identification of anchor items that are invariant across groups (i.e., from T1 to T2) in order to equate some assessment parameters across time. Anchor items were selected based on the four-

step process outlined by Woods (2009): (1) test all items for DIF using all items as anchors, (2) calculate the log likelihood for each item, (3) rank the log likelihood statistic from smallest to largest, (4) select the items with the smallest log likelihood. This process resulted in at least one anchor item per dimension within each rating source (i.e., task-related direct report ratings, socio-emotional direct report ratings, etc.). The final step in testing for DIF involves an iterative process of comparing item parameters over time until the same items are consistently identified as demonstrating significant changes between groups. The '*DIF'* function in the *mirt* package in R was used for this procedure using full information maximum-likelihood estimation.

Results from the above process support *IRT-H4* and time does contribute to significant shift in item parameters for select items across both dimensions and both rating sources (Table 7). Overall, DIF items demonstrate non-uniform crossing DIF in which the likelihood of raters choosing responses changes in location and discrimination as a result of time. Direct report ratings consistently show that ratings trend toward low-ability behaviors over time (Figures 8 and 10). Evidence of this trend is most obvious in the item trace lines for the socio-emotional item labeled 'q22' (Figure 10). Here, response options appear to shift more than a full standard deviation in leader ability where a rating of '4' in 2009 is equivalent to a rating of '3' in 2013 based on a 5-Likert scale.

Peer ratings resulted in the most items demonstrating DIF due to time compared to direct report ratings (Table 7). Again, DIF appears to be non-uniform crossing in all cases and shows an almost full SD shift in categories in the negative direction (Figures 7 and 9). The only item that does not follow this pattern is the task-related item labeled 'q15' where the shift in ratings appears to happen at the positive end of the response options (Figure 7). In 2009 the response categories 4 and 5 were able to differentiate among average and above-average leaders for q15;

however, in 2013 response options 4 and 5 are equally probable for slightly above average and high ability leaders. These results support *IRT-H4* and time does have an impact on both direct report and peer ratings for select MSF assessment items.

Part II – SRM Analyses

The goal of Part II of the study was to assess whether the round-robin MSF assessment data would provide interpretable variance components that could potentially provide more information regarding leader ratings on the two main leadership dimensions. Initial tests of the full SRM model including all variance components as random effects indicated that there was no variance attributable to group for task-related (.010, p= .511) or for socio-emotional ratings (.007, p= .678). Effects of dyadic reciprocity were also null for task-related (.004, p= .709) and for socio-emotional ratings (.012, p= .350). Comparisons based on change in log likelihood (Table 8) supported use of the constrained model for task-related (χ^2 (2)= .5, p= ns) and for socio-emotional ratings (χ^2 (2)= 1.0, p= ns) and as a result, these two effects were dropped from the model. These results are consistent with previous studies estimating SRM components from round-robin ratings of leadership (Kenny & Livi, 2009).

Hypothesis *SRM-H1* tested whether *perceiver, target,* and *relationship* components all account for statistically significant amounts of variance in leader ratings. Results of Wald tests on the variance components indicate that for task-related scores, *perceiver* (26%), *target* (46%) and *relationship* (28%) components all account for a significant amount of variance (Table 9). Similarly for socio-emotional scores, *perceiver* (27%), *target* (46%), and *relationship* (28%) components each account for a significant amount of variance (Table 9). Thus, *SRM-H1* is confirmed for both task-related and socio-emotional ratings.

The second set of hypotheses for Part II of the study estimate whether variance components differ for ratings on the two leadership dimensions. Hypothesis *SRM-H2* was supported and *target* variance (i.e., perceiver agreement about target leaders) was greater than *perceiver* and *relationship* variance for task-related ratings. Hypothesis *SRM-H3* predicted that *relationship* variance (i.e., variance related to the unique connection between the rater and leader) would be greater than *perceiver* and *target* variance for socio-emotional ratings. Surprisingly, the percentages of variance for the components were almost identical for both task-related and socio-emotional outcomes, but consistently there is more variance in the socio-emotional ratings. Therefore, *SRM-H3* was not supported.

Part III – Self-enhancement Analyses

The goal of Part III of the present study was to investigate a new way of measuring leader self-enhancement in MSF assessment ratings and determine whether this new self-enhancement measure is related to leader effectiveness outcomes. As discussed in the Leader Self-Enhancement section above, the present study uses round-robin MSF assessment data to calculate a measure of leader self-enhancement that has not been previously estimated with this type of data. This method of calculating leader self-enhancement provides a more refined measure of leader self-enhancement compared to the more conventional discrepancy scores typically considered in SOA analyses in MSF assessment ratings (Atwater & Yammarino, 1997) and in a previous measure that included a social comparison index (Kwan et al., 2004). All the parameters used to measure the impact that leader self-enhancement has on leader effectiveness outcomes are shown in Figure 11.

Two separate models of leader self-enhancement were estimated for task-related and socio-emotional ratings and served as the basis for testing the hypotheses in Part III of the

present study. The baseline model (Figure 11) implemented the same constraints as were used in the SRM from Part II with group variance as well as dyadic reciprocity constrained to be zero. Nested model comparisons of the baseline model were initially tested to determine whether the new model provided the best representation of self-enhancement in the present sample of leaders for both task-related and socio-emotional ratings. These nested models tested several possible scenarios against the baseline model for leader self-ratings: (1) self-perception is the same as other-perception, (2) the *perceiver* effect is the same as the self-rating, (3) the *target* effect is the same as the self-rating, (4) there is no self-enhancement in self-ratings, and (5) variance in enhancement of self is equal to the variance in the enhancement of others. None of the nested models provided a better fit to the task-related or socio-emotional models; therefore the originally hypothesized baseline model was used to test the remaining hypotheses.

Hypothesis testing: *SE-H1, SE-H2, SE-H3* and *SE-H4*. Hypothesis *SE-H1* tested whether leader self-enhancement, as measured through the SRM will negatively predict leader quality, as rated by the leader's direct reports. Results of this analysis (Table 10) reveal that leader self-enhancement positively predicts leadership quality for task-related ($\beta = .191$, $p = .002$) and for socio-emotional ($\beta = .211$, $p < .001$) ratings. In other words, leaders with higher self-enhancement receive higher leader quality ratings on average. These findings are opposite of the expected findings; thus *SE-H1* was not supported.

Hypothesis *SE-H2* tested whether leader self-enhancement would positively predict leader termination. Employee turnover status was converted into a continuous variable to meet the assumption of multivariate normality for use in the structural equation model for estimating leader self-enhancement. The new variable calculated the amount of time each leader worked at the organization before he or she was involuntarily terminated. Voluntary employee turnover

status, including Retired, Leave of Absence, Severance, and so on, were treated as having missing data. Finally, in order to include employees who were still active in the organization, a hypothetical *time of employment* variable was calculated assuming a scenario in which these leaders would spend the rest of their careers at the present organization and retire at the age of 75. Using this estimated amount of time an employee would work for the organization, based on a future retirement date for these active employees, allows for the comparison on a continuous scale of employees who have been terminated to those who are still with the organization. I addition, this approach resulted in a more normal distribution than if dichotomous dummy coding and been employed. In this new termination scale, lower values reflected a quicker rate of termination for individual leaders.

Results from testing of *SE-H2* (Table 10) indicate that leader task-related self-enhancement did not predict leader termination (β= -.255, *p*= .283); however, leader socio-emotional self-enhancement did predict leader termination (β= -.546, *p*= .007). Hypothesis *SE-H2* predicted that self-enhancement would predict quicker leader termination. Therefore, the significant, negative impact of leader socio-emotional self-enhancement on leader termination supports *SE-H2* for socio-emotional ratings that leaders who self-enhance would be more likely to get pushed out of the organization.

Hypothesis *SE-H3* tested whether leader self-enhancement would negatively predict leader promotions. Leader promotions were calculated by determining the number of levels each leader had been promoted to since taking the MSF assessment by the time data were pulled for the present study. Due to the relatively short amount of time since the assessment and data pull, the promotions variable resulted in a similar bi-modal distribution that was seen previously with employee termination data because employees were either in their same level with a score of 0 or

they had been promoted up one level with a score of 1. This result effectively created a dichotomous variable indicating whether or not a leader had been promoted. In order to create an interval measure, the level that each leader was in at the time of the data pull was used instead. There are four leader levels within the *middle tier* of leadership at the present organization and these were coded with a score of "1" indicating the lowest leader level and a score of "4" indicating the highest. Hypothesis *SE-H3* now tests whether self-enhancement is more common in leaders who are at higher levels within the *middle tier* at the present organization. Results from testing of *SE-H3* (Table 10) indicate that leader self-enhancement does not predict leader level for task-related (β= -0.068, *p*= .580) or for socio-emotional (β= -0.099, *p*= .322) ratings. Therefore, Hypothesis *SE-H3* was not supported.

Finally, hypothesis *SE-H4* tested whether leader self-enhancement was negatively related to leader performance. As stated earlier, the present organization rates employees separately on task-related and socio-emotional performance. The separate task-related and socio-emotional performance ratings were used to test the separate task-related and socio-emotional models, respectively. Results of this analysis (Table 10) reveal no relationship between task-related self-enhancement and task-related performance (β= -0.010, *p*= .122), or between socio-emotional self-enhancement and socio-emotional performance (β= -0.016, *p*= .799). Therefore, hypothesis *SE-H4* is not supported.

Discussion

The present study set out to answer three main questions related to use of MSF assessments in organizations: (1) how well does MSF assessment data inform leadership programs in terms of individual leader and relational processes from a psychometric perspective, (2) does the SRM provide more relevant information regarding relational processes of

leadership, and (3) can the SRM better inform leadership development plans by calling attention to issues relevant to leader self-regulation. Overall, the analysis results along with other findings that were not specifically hypothesized fully address each of the main questions posed above, providing information that is potentially valuable to organizations that use and interpret leadership ratings based on MSF assessments.

Part I – IRT

Item discrimination. Part I of the present study used IRT to evaluate MSF assessment item and test parameters from ratings gathered in a real organizational setting. The first set of hypotheses tested whether custom items written to specifically measure leader ability based on a tiered competency model were able to discriminate better for leaders within each of three tiers (i.e., *foundational, middle,* and *top tier).* Unfortunately, tests for *foundational tier* leaders were not possible because the GRM model did not fit the data. Small sample sizes within each rating category likely contributed to the lack of model fit.

For leaders in the *middle* and *top tier* competency groups, hypotheses were mostly supported with the exception of socio-emotional items across all rating groups for *top tier* leaders. However, compared to *middle tier* leaders, *top tier* leaders had more tailored items providing the highest discrimination values relative to core items across all rating groups for task-related MSF assessment items. This discrepancy in discrimination parameters between task-related and socio-emotional items for *top tier* leaders is likely due to the nature of responsibility at this level. At higher levels of the organization, leaders face increasingly ambiguous situations which are likely to impact socio-emotional ratings in one of two ways: (1) tailored items are less directly relevant for *top tier* leaders because specific leadership scenarios are more difficult to anticipate, therefore making it difficult to write appropriate items, and/or (2) ambiguity in socio-

emotional interactions with the leader make it more difficult for raters to observe specific behaviors. Task-related behaviors, on the other hand, may be more obvious to raters in the case of *top tier* leaders because these leaders have greater visibility in the organization and more people are directly impacted by the business success of these *top tier* leaders. In other words, the task-oriented behaviors of *top tier* leaders is directly related to performance of others in the organization, and so raters observe this type of behavior more keenly.

Item discrimination for *middle tier* leaders favored tailored items in every case except for direct report ratings of task-related items. This was surprising, since raters who report directly to *middle tier* leaders presumably have the most exposure to the leader's behaviors, and therefore would be better observers of behavior. One possible explanation for direct reports providing less discrimination for tailored task-related items is a general lack of understanding regarding tasks that leaders perform at this level. Compared to *top tier* leaders, these leaders are more likely to have direct reports who are not also considered leaders in the organization, and therefore have had less exposure to the competency model tailored to this leadership level. These lower-level employees are more familiar with the *foundational* leader behaviors, which may explain why they provide more discriminatory ratings for these types of behaviors. Socio-emotional behaviors on the other hand, may be easier for lower-level direct reports to discern since these behaviors directly impact the relationship they have with the leader.

Results from this section suggest that different types of raters do provide unique pieces of information regarding leader behavior. Furthermore, the use of both core and tailored items with MSF assessments appears to be a good strategy to cover a broad range of leader behaviors from multiple perspectives.

Change over time. The second set of hypotheses in Part I addressed changes in test and item parameters over time for *middle tier* MSF assessments. Results indicated that test information did change over a period of four years, but not in the hypothesized direction in most cases. The hypotheses in this section assumed that test information would degrade over time as raters become more familiar with competency behaviors. Test information for task-related items did shift in the hypothesized direction for peer ratings on the task-related dimension, but ratings from direct reports showed the opposite pattern, providing more information in 2013 compared to 2009. The different shifts in test information over time could be a result of how long the raters have known the focal leader. Peers could presumably have a longer history observing behavior of the leader, which could distort their observations over time as behaviors become more normalized, as was originally hypothesized. Direct reports on the other hand may have known the focal leader for a shorter period of time, and therefore their ratings of behavior provide more of a snapshot in time as opposed to a trend over time.

The socio-emotional dimension demonstrates a less clear pattern of test information for both peer and direct report ratings over time. Direct report ratings demonstrate a similar pattern to task-related dimension information, with 2013 ratings providing more information than 2009 ratings. However, peer ratings of the socio-emotional dimensions appear to provide the same information over time, but the test provides more information for lower ability leaders in 2013 compared to 2009. This pattern of test information suggests that item parameters are likely changing over time for peer ratings of socio-emotional items, which was investigated more thoroughly in the final set of hypotheses for Part I.

The final hypothesis in Part I (*IRT-H4*) tested whether item parameters for leaders who re-take the MSF assessment change over time. Results from tests of differential item functioning

(DIF) indicated that hypothesis *IRT-H4* was supported and some item parameters do change as a function of time. In all cases, items demonstrated crossing, non-uniform DIF that was relatively consistent across items. Items shifted in the negative direction providing more information for low-ability leaders such that a rating of '2' in 2009 was equivalent to a rating of '1' in 2013, but ratings provided better discrimination for above-average and for high ability leaders in 2009. Item responses also show varying probabilities in the likelihood of selecting each response category based on leader ability over time; for low-ability leaders the probability of response options generally went up, but for high-ability leaders the probability of a rater selecting a positive response option at the high end of the response scale generally went down.

Results suggest also that DIF impacts item parameters regardless of item type with an almost equal number of core and tailored items demonstrating DIF. Furthermore, task-related peer ratings had the most DIF and socio-emotional peer ratings had similar levels of DIF compared to direct report ratings over time. One possible explanation for this trend is that as leaders and their peers progress in their respective careers, their perceptions of leader ability are changing. As hypothesized, leaders' exposure to the same competency model and observing the same types of behaviors (i.e., leadership behaviors) over time is suggesting a shift in how raters use the response scale. Another possible explanation is that peers are more likely to become close friends with each other after working together for several years, which makes them less likely to provide honest feedback for fear of jeopardizing a social relationship. This could explain why, over time, items with DIF tend to provide more discrimination at lower levels of leader ability.

Finally, a few items seemed to demonstrate an extreme amount of DIF that are worthy of note. For peer ratings, socio-emotional Item 2 (Figure 9; q22) demonstrated DIF such that in

2013 this item provides no discrimination for leaders who are above average because the response option '5' has the highest probability of being selected for leaders of average ability. The content of this item is related to diversity and whether a leader demonstrates sensitivity to cultural differences of others they work with. The parameter shift of this item suggests that a diversity item may not provide the best differentiation among leaders. Another item of interest is from the peer ratings of task-related Item15 (Figure 7; q15). This item is the only case of a response category increasing in probability for high-ability leaders in 2013 as compared to 2009. The content of this item asks about the leader's perseverance in pursuit of his/her goals. In this case, peer raters may be better able to discern whether another leader is achieving his or her goals after they have been working together for a period of four years. This could explain why response options for this item would generally be higher in 2013 than in 2009 for high-ability leaders.

An interesting finding from these analyses is that the MSF assessment items appear to differentiate best among low- and average-ability leaders but do not provide information that can be used to discriminate among above-average leaders, a phenomenon that appears to be more pronounced over time. A likely explanation for this is how raters are using the response scale. For ratings of most leaders, raters only utilize the higher points on the scale. The lower scale points are reserved for leaders who are drastically worse than other leaders are on average. Therefore, more variability in responses only exists for leaders who are below average in ability.

Part I – Practical Implications

Results from Part I of the present study provide sufficient empirical evidence to answer the first research question: How well does MSF data inform leadership programs from a psychometric perspective? Results suggest that MSF data provides valuable information but do

not do very well as far as informing about leadership development over time. First, results support the use of both core and tailored items on MSF assessments. In most cases, tailored results provided better discrimination at various levels of leadership, but core items also played a role. Furthermore, there is evidence that different types of raters have distinct perceptions regarding core and tailored items, so using a broad range of items maximizes information gathered from the assessment. This is especially relevant to organizations that employ a tiered leadership competency model where specific behaviors may be more or less salient to employees at different levels of the organization.

An important practical implication of findings from Part I of the present study is regarding change in individual leader scores over time. Leaders who may be diligently tracking MSF scores over time may be receiving flawed, or even false, information that hinders their ability to effectively absorb feedback and apply this information in forms of feedforward control that are necessary when their leadership responsibilities increase. In other words, leaders may receive conflicting information regarding assessment results that could negatively impact their ability to develop task-related and socio-emotional skills as they progress in their careers and encounter more ambiguous leadership situations.

Aside from the occurrence of shifting item parameters, their low discrimination for average and above-average ability leaders suggests that scores from these assessments may provide these higher ability leaders with almost no useful information. As stated previously, some of the item characteristic curves suggest that there are no distinctions among ratings of ‘3’, ‘4’, or ‘5’ on the 5-point Likert scale used for distinguishing among high ability leaders. This creates somewhat of an ethical dilemma for organizations who may put pressure on all leaders to improve scores over time. If organizations do not track the psychometric properties of MSF

assessments over time, they may be actively providing leaders with misinformation regarding their task-related and socio-emotional ability.

One practical recommendation would be to write leader MSF assessment items to specifically cover a broad range of leader behavior rather than writing neutral items that cover specific content areas as was done in the present organization. In other words, leaders at all levels would be rated on the same items that are written to address the broadest range of leader behaviors that reflect the ability of the leader within the broader leader population. These items should be written to specifically ask about particular actions or experiences the leader has had rather than general questions about his or her behavior. As leaders master various competencies, the likelihood of raters endorsing more items of this type should go up over time, which may provide two important pieces of information for the organization: (1) more consistent tracking of individual leader progress, and (2) a general profile of leadership based on which items are endorsed for any particular leader. These leader profiles may allow for more nuanced leadership tiers demonstrating leader competence in more finite areas of each dimension as opposed to focusing on aggregated dimension scores.

Part II – Social Relations Model

The goal of Part II of the present study was to understand whether conventional MSF assessment data could be reconfigured into a SRM framework in order to provide more relevant information regarding relational processes of leadership. Hypothesis *SRM-H1* was supported, with each variance component of the SRM contributing significant amounts of variance to ratings of task-related and socio-emotional leader ability. However, variance components only demonstrated the patterns that were hypothesized in *SRM-H2* but not those hypothesized in *SRM-H3.* As predicted, *target* variance explained the most variability in task-related ratings, in

support of *SRM-H2*; however, *relationship* variance did not explain the most variability in socio-emotional ratings, and so *SRM-H3* was not supported. In fact, the variance components showed essentially the same pattern across task-related and socio-emotional ratings, with the majority of variability coming from *target* variance, the second highest source of variability coming from *perceiver* variance (i.e., systematic patterns in how peers tend to rate others), and the lowest amount coming from *relationship* variance (i.e., rating variability attributable to the unique relationship between the leader and the rater).

Although inconsistent with what was hypothesized, the presence of variance components of relatively the same strength on the two dimensions of leadership are a good sign that raters at the present organization are taking the MSF assessment seriously. Additionally, the variance components reported in the present study mirror those reported by previous researchers, with the exception of *perceiver* variance. In their review of round-robin ratings of leadership, Kenny and Livi (2009) found that stable *target* variance of leadership ratings ranged from a low of 30% to a high of 60%, which are similar to the *target* variances reported above. However, Kenny and Livi (2009) found much smaller amounts of *perceiver* variance, ranging from zero to 14%, compared to those in the present study (26-27%). Though the present study similarly analyzed round-robin ratings of leadership, the studies reviewed by Kenny and Livi (2009) did not use real organizational ratings of leader ability, and this may explain the higher relative *perceiver* variance reported in the present study.

Despite consistent findings in *target* variance across both dimensions of leadership ratings in the present study, at best only 44% of variability in leader ratings was attributed to how leaders were consistently perceived by others. In other words, more than half of the remaining variability in leader ratings can be attributed to: (1) systematic response patterns of particular

raters, (2) the unique relationship between the leader and the rater, and (3) error. These results speak to the potential issues of using aggregated leader scores to differentiate among leaders. Therefore, aggregated MSF assessment scoring as it is conventionally performed may not be the most appropriate tool for estimating the overall ability of the leader.

As mentioned previously, SRM analyses controlled for gender and tenure. Tenure did not demonstrate a significant effect on MSF ratings, but there were some interesting effects of gender. As reported in Table 9, there was a significant main effect of the perceiver's gender in rating variability of task-related outcomes such that women raters tend to provide higher ratings overall. Furthermore, the interaction between target and perceiver gender was also significant for both task-related and socio-emotional ratings. These results indicated that ratings are significantly higher when women rate other women compared to any other gender combination (i.e., men rating women, men rating men, etc.). In other words, men provide similar ratings regardless of the gender of the leader they are rating, but female ratings vary depending on the gender of the leader they are rating. Previous researchers have reported significant gender interactions with performance ratings provided by peers (Pulakos, White, Oppler, & Borman, 1989); however, these types of effects have been explored less often with traditional MSF assessments. Moreover, the results from the present study do not explain whether ratings made by men or women are more accurate, making further interpretation of the rating interaction difficult. For example, the present results could be interpreted to suggest that women tend to rate higher than men on average, and this effect is even stronger when they rate other women. It may also be the case that men under-rate other leaders on average, while women rate other leaders more accurately.

Part II – Practical Implications

Results from Part II of the present study provide sufficient empirical evidence to answer the second research question: Does the SRM provide more relevant information regarding relational processes of leadership? Organizations would be best served by analyzing their data in a relational fashion to understand what is most contributing to leader ratings before they use a MSF assessment score as a metric in organizational decision-making or for leader development activities. The significant amount of *perceiver* and *relationship* variance in ratings of leader behavior suggest that round-robin data collection of MSF assessment data may provide an organization with more useful information regarding leader behavior, as well as the extent that social processes in the organization are impacting ratings of leadership.

In the present organization, *target* variance explained the most relative variability in ratings compared to *relationship* or *perceiver* variance. However, in other studies of leadership ratings using the SRM, *perceiver* variance was found to predict a majority of variability in ratings (Kenny & Livi, 2009). If this were to be the case at another organization, MSF assessment ratings should be given much less weight when used for personnel decision-making than other potentially more concrete measures such as simulations or assessment center data. More research is needed to understand whether SRM components alone can provide a valuable metric for leader assessment, but nonetheless results from Part II of the present study suggest that the SRM would serve as a valuable complementary analysis to more conventional MSF analysis strategies.

Another relevant practical finding from Part II of the present study is that effects due to rater may be even more profound than G Theory studies of MSF ratings have previously suggested (Greguras & Robie, 1998; Mount et al., 1998). The analysis of peer ratings in the SRM indicates that a substantial amount of variance in ratings is due to individual rater effects or

idiosyncratic relationship effects, which is consistent with effects reported by Greguras and Robie (1998). Furthermore, these influences may be even larger in ratings from direct reports or managers because these rating sources have additional power distance issues involved in the relationship between the leader and the rater as well. Thus, organizations may be well served in providing leaders with some indication of primary sources of variability regarding their MSF assessment data. This way, leaders can more accurately interpret assessment feedback.

One way for organizations to inform leaders about rating variability would be to train feedback coaches on ways to present this information to leaders that would be beneficial to individual development plans. Leaders with high rating variability might be encouraged to conduct focus groups with their peers or subordinates in order to gain a better understanding of where their discrepancy in ratings is coming from. This type of feedback could be highly valuable to leaders in understanding how their behavior influences the broader social system. An opportunity to receive face-to-face feedback could also promote a leader's understanding of how their behavior influences the broader organizational system; however, organizations would need to take care to minimize potential negative repercussions for this type of activity.

Leadership researchers have started calling for more research focus on the influence that follower perceptions have on leadership processes and what the useful information this provides in support of leader development activities (Harms & Spain, 2014; Lord & Dinh, 2014). The present study suggests that these same perceiver effects are also worth considering for peer ratings.

Part III – Self-Enhancement

Part III of the present study investigated whether a measure of self-enhancement that combines both Self Insight Theory (Allport, 1937) and Social Comparison Theory (Festinger,

1954) is able to predict leader characteristics. As mentioned previously in the Theories of Self-Perception section, conventional MSF assessment analyses utilize only Self Insight Theory to measure leader self-awareness. This model tests self-other agreement (SOA), and leaders who rate themselves higher than others (i.e., over estimators) are assumed to have worse leadership outcomes than leaders who agree with other's ratings (i.e., agree high/agree low) or rate themselves lower than others (i.e., under estimators). However, evidence provided above suggests that the SOA does not consistently predict leader outcomes, and these inconsistencies are likely related to leader social comparisons that are not being measured in SOA. Social comparisons may contribute to another aspect of self-serving bias (SSB), or inflation of self-ratings. Including the focal leaders' ratings of others provides valuable information about how the leader sees him/her self compared to other leaders. Overall, it was expected that leaders who scored higher on self-enhancement, as calculated by the SRM, would experience more consistent and extreme negative outcomes such as those predicted by over estimation in the SOA framework.

The initial test of the self-enhancement model indicated that overall, leaders at the present organization demonstrate relatively little self-enhancement in both task-related and socio-emotional ratings. However, despite the low overall estimate of self-enhancement, the baseline model, which assumes significant self-enhancement among leaders, provided the best fit to the data compared to other models. This finding indicates two important characteristics of the self-enhancement data. First, the baseline model tested in the present study indicated that a significant amount of variance was attributable to self-enhancement, but the effect was small; this is an indication that there is a high level of variability among leaders regarding whether or not they enhance self-ratings. Second, variability in the amount of self-enhancement that leaders engage

in suggests that self-enhancement is likely related to other leadership effectiveness outcomes such as leader quality, turnover, leader level and performance.

Although results from Part III provide evidence that leader self-enhancement is related to leader effectiveness outcomes, the direction of these relationships suggests a much more complicated relationship between leader self-enhancement and leadership outcomes than originally hypothesized. Contrary to what was hypothesized in *SE-H1*, leader self-enhancement of both task-related and socio-emotional ratings was positively related to leader quality ratings. The only other significant association between leader self-enhancement and leader characteristics was a significant relationship between socio-emotional self-enhancement and leader termination, which was in the hypothesized direction. Together these results provide an interesting perspective on how leader self-enhancement impacts leadership effectiveness.

A finding that was not expected was that self-enhancement, which provides a much more refined estimation of SSB in leader self-ratings, is strongly associated with positive leader outcomes. Contrary to MSF assessment findings regarding leader over-estimation, findings from the present study suggest that leaders with an inflated sense of self as compared to their peers are rated more positively by direct reports on a measure of leader quality. This finding might suggest that leaders who inflate self-ratings are demonstrating superior leadership behaviors that are being detected by direct reports, and are not being recognized via peer ratings. However, results from tests of hypothesis *SE-H2* suggest that leader self-enhancement on socio-emotional ratings are associated with leader termination. So while self-enhancement may be associated with seemingly effective leadership, these leaders may still have ultimately negative outcomes within the organization, especially when they self-enhance on ratings of socio-emotional leader behaviors. Together, these findings in Part III of this study provide strong evidence that leaders

who self-enhance are the leaders who are most disconnected with the reality of their leadership ability.

Leader level and leader performance were not associated with leader self-enhancement. One possible reason for the null relationship between self-enhancement and leader level could be the relatively short amount of time between when leaders participated in MSF assessments and when level data was collected. MSF assessment data were collected in 2013 and 2014 while leader outcome data was collected in 2015. It is possible that promotion rates might be more strongly related to leader self-enhancement over a longer period of time. Had there been more time between the assessment and when promotion data was collected, results may have demonstrated a pattern between leader self-enhancement and individual promotion rates.

Leader performance, on the other hand, may have suffered from the opposite issue. Leader performance ratings used in the present study were from the year 2015. In some cases, this was two years after the leader participated in the MSF assessment. It is possible that the relationship between leader self-enhancement would be most pronounced when measured in the same year. The present organization participates in employee feedback and leaders who rated high on self-enhancement in 2013 or 2014 may have participated in coaching sessions or other development activities to mitigate the potentially negative impact self-enhancement tendencies may have on performance outcomes. Nevertheless, results suggest that there is no association between leader self-enhancement and future performance, as was hypothesized.

The present study specifically hypothesized about the relationship between leader self-enhancement and various leadership outcomes. As stated in the Leader Self-Enhancement section above, the present measure of leader self-enhancement accounted for the impact of both *k* (perceiver effect) and *q* (target effect) on leader self-ratings. Results in Tables 10 - 13 reveal a

systematic pattern in how *k* and *q* influence leader self-ratings. First, *q* significantly predicts leader self-ratings for both task-related and socio-emotional ratings. However, the influence of *k* on leader self-ratings shows a slightly different pattern with *k* significantly predicting variability in leader socio-emotional self-ratings, but not significantly predicting variability in leader task-related self-ratings. In other words, there is no systematic variance in leader task-related self-ratings attributable to how the leader rates others.

Other relevant post hoc findings concern the influence of *perceiver* and *target* effects on leader effectiveness outcomes in Tables 10-13. Interestingly the *target* effect (i.e., ways that raters view the focal leader) did not predict any leader effectiveness outcomes. In other words, attributes of the leader were not related to leader effectiveness. However, in most cases the *perceiver* effect did predict leader promotions, performance, and turnover. This means that how leaders tended to rate other leaders had a strong influence on their effectiveness outcomes, or conversely, these results could be interpreted to suggest that when leaders were more effective they tend to rate fellow leaders higher. This pattern of effects was surprising, and provides convergent support for the possibility that rater effects have a substantial influence on leader ratings. At the very least, results presented above suggest that leader effectiveness as well as the way a leader views him or her self have a relatively strong impact on the way a leader rates others. These rater effects likely contribute to variability in ratings in addition to other well-known rater effects such as rater type and rater motivation.

Part III – Practical Implications

Results of Part III provide sufficient empirical evidence to answer the research question: Can the SRM inform leadership development plans by calling attention to issues relevant to leader self-regulation? Self-regulation is important here because it concerns the process through

which leaders are motivated to change their behavior or engage in development activities based on feedback gathered from a MSF assessment. Results strongly suggest that leader ability is only part of the story, and that social processes underlying leadership are just as relevant to developing both leader and leadership abilities. Organizations wishing to use MSF ratings to differentiate among leaders and to potentially identify leaders who will be successful at higher levels in the organization will need to consider determinants of leader ability beyond inherent skill competency; namely, whether or not their organization has appropriate social system to support leader development.

Incorporating round-robin data in the analysis of leader self-ratings expands upon the benefits of the SRM in ways that are even more relevant to leader self-regulation activities than those previously mentioned in Part II. This model not only provides leaders with some insight regarding their own rating tendencies, but also an indication of how difficult it is for raters to predict ambiguous leader skills, particularly socio-emotional skills.

The SRM estimates of self-enhancement may be most relevant to socio-emotional leader behaviors, a finding that is consistent with the idea that social processes of leadership are more relevant in ratings of more ambiguous leader behaviors. In the present study, leaders with high socio-emotional self-enhancement were more likely to be terminated in a short amount of time compared to leaders who had lower levels of socio-emotional self-enhancement (Table 10). High levels of self-enhancement may serve a leader well with task-related behaviors; however, this biased view of the self, especially when it does not factor in socio-emotional ability, may result in negative organizational consequences for the leader. Leaders with grandiose ideas about their own leadership are not likely to become truly engaged in development activities, and this may impede them the most when it comes to the so-called “softer skills” related to leadership. This

impact is likely to manifest in negative outcomes for high-level leaders because these leaders are likely to rely more and more on socio-emotional skills they do not have as they assume more responsibility within the organization. Leaders who are unable to motivate and influence their followers are likely to perform poorly in these new roles.

An organization may want to monitor the amount of self-enhancement in the leadership population and track other leader effectiveness outcomes that may also be related to leader self-enhancement. This could provide an organization with powerful tools on how to better set leaders up for success as they rise through their specific organization.

Study Limitations

The present study investigates the distinction between leaders and the process of leadership as well as the important role that MSF assessment data plays in shaping our understanding of what these terms mean in organizational settings. However, this study was not immune to some limitations that are relevant to the interpretation and generalizability of the results presented above.

First, the present study used data from a single organization. Analyzing data from a single organization limits the generalizability of the results presented here. Even though it is possible that MSF assessment data from another organization will demonstrate different psychometric patterns and different relative variance components than were found at the present organization, the present results still indicate that these analysis strategies would be useful to practice for any organization. Furthermore, using data from a single organization allowed for the tracking of assessment data over time among a consistent pool of leaders and raters, something that was necessary to do in order to investigate how rater perceptions might change over time. The single organizational context also afforded the opportunity to reconfigure MSF assessment data into a

round-robin format to calculate components of the SRM. Even though the exact estimates and relationships may not be replicated with data from another organization, the implications for what these different effects mean does not change. In other words, social dynamics in leadership are likely to impact leader ratings in any context even though the profile of the relative variance components is likely to change in a different population of leaders.

Another potential limitation related to using data from a single organization is the historical context during which these data were collected. Analyses conducted from 2009 to 2013 were likely impacted by a large organizational transformation effort that has slowly been evolving at the present organization over several years. As stated previously, large-scale change could impact psychometric properties of assessment items because communication and messaging as part of the transformation may alter the ways in which employees view both leadership and leader behaviors. These changes may then influence the way raters react to assessment items in ways that could not be measured with the present study design; however, all leaders in the present dataset were presumably impacted by the change equally since they had all been at the organization between 2009 and 2013. Therefore, organizations that do not undergo transformation efforts may not see the same amount of DIF in assessment items over time as was found in the present organization, but for the results presented here it is assumed that leaders were uniformly impacted by these potential changes in organizational processes.

Another limitation of the present study involves the process for compiling the round-robin ratings of MSF assessment data. As stated previously, these data were originally collected as a one-with-many design whereby several raters provide ratings about one focal leader. Reconfiguring this dataset resulted in the loss of a majority of the MSF assessment ratings from 2013 and 2014. Even though leaders retained in the round-robin dataset demonstrated similar

demographic and dimension scores to the overall leader population, there may be some social influences among these leaders that does not exist among those leaders who were excluded. In other words, variance components may be slightly higher than they would be in a broader leadership population because leaders included in the present round-robin dataset consist of those leaders who purposefully selected the same peers to rate each other. Although this may contribute to a subtle influence on SRM parameters, the possibility that unique social processes exist among these leaders is nonetheless not expected to change the results dramatically because the leaders did not know in advance that ratings would be used in this fashion. In other words, leaders were not actively picking peers who they knew would also be rating them in a round-robin context. However, if the SRM is openly used in future applications of MSF, it is possible that such an effect would warrant concern.

Another limitation related to challenges in configuring the round-robin dataset involves missing data. Since leaders were not purposefully told to rate each other in a round-robin fashion, most groups have at least one dyadic pair lacking a rating. In many other groups, there may be one leader who has provided ratings for everyone else but no one else in the group has provided ratings for that same leader. Despite the missing data, it was possible to conduct the analyses without needing to make assumptions about imputing missing data values, which bolsters the credibility of the analyses. However, future studies should aim for more complete round-robin rating groups in order to maximize the power of predictive relationships between variance components and leader effectiveness outcomes.

Future Research

The groundbreaking nature of the present research opens up a number of potentially useful directions for future research. First and foremost, a line of research could further delineate

the distinction between the qualities of a leader and the qualities of the leadership process, an effort that is likely to require a series of studies on leadership. Findings from the present study also provide empirical support for conducting studies on a wider range of factors that may contribute to perceptions of leaders within an organization. More specifically, factors associated with providing more accurate developmental feedback to leaders based on self-ratings can be focused on findings in the present study that leader effectiveness ratings may impact how the leader views him or herself and how they view others. These perceptions have the potential to provide important implications for differentiating among leaders and developing leaders, and also for measuring leadership effectiveness.

Future psychometric research. As mentioned previously, researchers who focus on leadership assessments would be well served to investigate alternative item types that could maximize differentiation among leaders over time. An assessment like this would require identifying the appropriate range of leader behaviors that could serve a single organization through a leader's entire career – a considerable research undertaking. Ideally, the same set of questions could be used with employees at any level within the organization, with the items that are endorsed for each individual providing an indication of his or her leadership ability or leadership potential. Once an appropriate list of items is identified, longitudinal validation studies would be necessary to determine whether these types of assessments have predictive validity of future leader success, and maintain psychometric integrity over time.

Future research could also look further into some of the gender effects that were uncovered in the present study. Psychometrically, future research should consider whether there are gender differences in differential item functioning (DIF) over time. Previous work has investigated whether gender impacts ratings of leaders and found little impact of gender on item

parameters (Penny, 2010), but to the author's knowledge no one has considered how gender effects might change over time. The present study reported two findings that suggest a possible interaction effect of gender and time on item parameters: (1) one of the two items with a particularly large amount of DIF had asked about leader diversity awareness, and (2) gender predicted significant variance in leader ratings indicating that women rate women leaders higher than any other gender combination for both task-related and socio-emotional ratings (or conversely, that men systematically underrate all leaders regardless of gender). Together, these findings suggest that there may be fundamental differences in how men and women not only perceive leaders but also in how they experience leadership. It is also possible that over time, diversity initiatives within an organization may change the way certain groups perceive others as leaders. There is no way in the present study to fully explain the interaction effect of gender on variability in leadership ratings. Therefore, future studies could aim to understand: (1) whether women tend to systematically overrate other women or men tend to systematically underrate all leaders regardless of gender, (2) whether gender influences the change in item parameters of leader ratings over time in idiosyncratic ways, and (3) whether diversity initiatives have an impact on how majority and/or minority groups perceive leaders and the process of leadership in the context of leader ratings.

Relational processes of leadership. Another area ripe for future research involves relational processes of leadership. The present study provides strong support for organizations to utilize analytic methods with leadership data that reflect the level of theory and measurement that is most relevant to questions regarding the process of individual leadership. One of the most surprising findings in the present study was that *perceiver* variance was the strongest and most consistent predictor of leadership effectiveness outcomes. As stated previously, to the author's

knowledge the present study is the first to use organizational MSF assessment data in an SRM analysis effort, and so additional future research is first needed to determine whether this is a phenomenon of leadership in general or if these effects are unique to leaders in the present organization. As noted earlier, MSF assessment data is not a consistent predictor of several leadership effectiveness outcomes; therefore, it is possible that the significant perceiver effect reported here holds a clue as to why direct ratings of leader behavior do not systematically predict leadership outcomes.

It would also be useful in future research efforts to investigate other leader characteristics or leadership contexts to see whether these have an impact on leader rating patterns. For example, leader personality could interact in the SRM such that certain types of leaders tend to over- or under-rate other types of leaders. One possible explanation for the finding that *perceiver* effects predict leadership outcome variables is that there is a general agreeableness factor influencing leader ratings. In other words, if a leader is more agreeable and pleasant he or she may be more likely to rate others in a positive way, and this positive outlook on the behavior of others is what contributes to leader success. Another possibility is that social intelligence plays a role in how leaders rate other people that is also connected to their success as a leader. These leaders may consciously or subconsciously use their ratings of others in a way that strategically places them in a favorable position in their respective social network. This type of strategic positioning may also have a direct impact on other ratings of leader effectiveness or success.

Situational factors of the leadership context may also play a role in how rating patterns impact leader success. For example, leaders who work in closer proximity to others may be more sensitive to the way they rate others on MSF assessments and this sensitivity may be more easily picked up by other raters and reflected in this leader's effectiveness outcomes. Also, leaders who

work in functions that are required to be more collaborative with other activities or groups in the organization, such as human resources, may experience relational influences on leadership ratings in a different and possibly advantageous way. The present findings do not indicate whether these surprising findings are a result of individual leader differences or of the leader's overall context. Therefore, future research investigating possible organizational design factors that impact leader perceptions of others and leader effectiveness outcomes is definitely needed.

Leader self-enhancement. Gaining a better understanding of the determinants of self-ratings of leadership is another area that holds many opportunities for future research. The fact that rater influence may be larger than previously estimated and that leader rating tendencies appear to be related to leader effectiveness outcomes suggests that more research is needed to understand whether self-perceptions of leadership are actually beneficial in the context of leader development. The present study provides some evidence that addressing any tendency for leader self-enhancement may actually distract leaders from more concrete data that could better inform their goal-setting behaviors, especially when the current assessment techniques do not provide adequate discrimination over the full range of leader ability. Studies of complex organizational systems suggest that managerial understanding of the human network is crucial to reducing variability in organizational performance (Smith, 2015). In other words, leaders need clear feedback from their social network in order to successfully align their behaviors to be maximally effective in the organization. Therefore, the subtle nuances of leader self-ratings and circumstances under which SSB occurs in leader self-ratings need to be better understood.

One seemingly crucial element to understand leader self-enhancement more fully is the role that self-confidence plays in this relationship. A central argument in the present study was that not all leaders who rate themselves at the top of the rating scale are inflating their ability,

even when other raters do not rate them as high. Some of these instances of "over-estimating leaders" may be leaders who are accurately reflecting a high level of confidence in their own ability (Hollenbeck & Hall, 2004; Shipman & Mumford, 2011), which, contrary to potentially negative findings associated with over estimation, may be predictive of more positive leader outcomes for them. Even though some research suggests that inflating one's self-ratings could, in extreme outlier cases, be an indicator of so-called "dark personality" (Cullen, Gentry, & Yammarino, 2015), presumably, this type of leader would have poor self-regulation ability, which is then likely to have a negative impact on their overall effectiveness. In other words, leader effectiveness outcomes associated with over-estimation are likely to depend on whether leader self-ratings indicate an appropriate level of self-confidence.

Many large organizations, including the organization in the present study, identify leaders who they consider to be high potential individuals (Slizer & Church, 2010) and this variable could offer some explanation as to why self-enhancement may benefit leaders under some circumstances but result in negative consequences in other circumstances. Unable to consider their behavior in the context of a broader system, leaders with unrealistic views of their leadership ability would be unable to develop effective relational processes (Lord, Brown, & Freiberg, 1999; Carver & Scheier, 1982). Leaders who are self-enhancing in these detrimental ways may be more likely to experience negative career impact (Cullen et al., 2015), experience lower organizational commitment as well as lower quality relationships with subordinates. In contrast, managers who demonstrate self-enhancement that is in line with their ability or potential as a leader would be more likely to have built strong alliances, and would have benefited from self-regulation reinforced through years of constructive assessment of self and success in higher leadership roles, which in turn is likely to positively impact manager and

subordinate outcomes. Therefore, future research on self-enhancement would benefit from considering the potential moderating effect of employee high potential status or another measure of employee confidence on the relationship between leader self-enhancement and effectiveness outcomes.

Practical Implications for Organizations

The present study utilized existing MSF assessment data to explore whether there is value in considering MSF data from an SRM perspective. Results described above suggest that this type of model does indeed offer unique and potentially valuable information to leaders and organizations, especially when factors related to the social processes of leadership are the focus.

Improving existing MSF processes. Results from the present study have research-to-practice implications that are relevant to the ongoing debate about whether MSF data should be used for the dual purposes of leader development and evaluation in organizations. First, although results from Part I of the present study suggest that changes in scores over time may not be providing leaders with accurate information, this was only true for a subset of items. Therefore, an organization wishing to improve their MSF process would first need to perform their own psychometric evaluation to determine which MSF items do not vary over time. Once identified, stable items could be more confidently incorporated into performance appraisal metrics.

In contrast, items that vary over time could be repurposed to focus exclusively on leader development activities, especially if the concept of development shifted from the idea of tracking scores over long time periods to establishing a culture that provides opportunities for more immediate or real-time feedback. Recent perspectives on MSF development processes support the utility of real-time leader feedback because leaders are more likely to be receptive to feedback when they are the ones who initiate the process of generating this feedback. Rather than

implementing a formal and regular assessment process, organizations could promote self regulation in leader development by utilizing new technologies to implement a leader-initiated, real-time feedback process. Organizations could set up a system that allows leaders the opportunity to self-select specific items from a set of competencies that have been predetermined to be most relevant to their group. This semi-customized set of items could then be deployed to raters in a similar fashion within existing MSF processes. The importance differences here are that the leader self-initiates this process, and the resulting feedback is tailored to the leader's salient needs within their group.

Separating the traditional MSF assessment based on item type would also serve to maximize the benefit of an existing MSF process. Furthermore, if changes to the assessment and feedback process are communicated effectively, rating bias for development-specific items could be reduced for both self- and other-ratings, which would provide a leader with more accurate information regarding rater perceptions of his/her behavior. Furthermore, leader wellbeing may also be improved because a system that is able to separate development from performance appraisal is likely to alleviate potential anxiety leaders may have regarding other-ratings. In other words, if leaders are afforded the level of control they need to self-regulate the development process, they are likely to be more engaged and invested in the outcomes.

Incorporating a Social Relations Model perspective. Practically speaking, organizations interested in using the SRM analysis framework to find out how rating variability influences leadership effectiveness in their own leader population have two main choices, either: (1) restructure existing MSF assessment data into a round-robin format as was done in the present study, or (2) redesign MSF assessment data collection to include round-robin ratings of leaders. Large organizations with a regular MSF process and relatively consistent pool of

leaders may wish to explore SRM components in their existing data before formally incorporating these types of analyses into their overall process. On the other hand, smaller organizations may not have enough MSF assessment data to create a round-robin dataset with a sufficient number of groups for the SRM analysis.

Organizations wishing to incorporate round-robin data collection into their MSF assessment process may first choose to conduct a pilot study with leaders using a subset of their existing MSF assessment items. Previous studies of SRM components in leadership ratings (Livi & Kenny, 2009) were primarily lab studies where ad hoc teams were asked to rate whether or not members of a team acted as leaders during the designated lab activity. These types of studies included much fewer than the 351 groups analyzed in the present study, but were still able to estimate all variance components of the SRM because there is typically very little missing data in more purposeful round-robin data collection. An advantage of performing a controlled pilot study is that organizations can get initial estimates of what variance components look like within their organization before rolling this type of assessment out to the entire organization.

A pilot study would also provide the organization with an opportunity to gather qualitative feedback from leaders on how they view a round-robin process as opposed to relying on the more traditional one-with-many rating design. Reactions from leaders may become useful for organizations because a transition to round-robin data collection will likely require additional training and communications to educate leaders on the value of providing ratings on leaders who also rate them. Careful messaging regarding the added value of this rating strategy will increase the likelihood that leaders embrace the new process and take ratings seriously. Developing the right culture around round-robin leadership ratings may even serve to increase collaboration among leaders because these types of ratings emphasize the leadership context and relationships

as opposed to individual ability. Leaders who believe their organizations are considering the social context of their leadership behaviors may also be more willing to make themselves vulnerable by being sincere in self and other ratings, which could increase the effectiveness of the leadership development process in general.

The analyses that were conducted in this study can also be applied to organizational initiatives other than leadership. An obvious application would be for organizations to incorporate an SRM framework into team building activities. Variability in ratings of teamwork could provide valuable information regarding team processes that may benefit overall team performance. Applied to teamwork, the model used in the present study could provide team members with valuable feedback on their working process that could contribute to self-regulation activities such as goal setting for the team as a whole. For example, team member self-enhancement may have a very different impact on how team members rate each other than was found with peer leaders in the present study. Additionally, there may be much stronger perceiver effects among team members because individual performance has a much bigger impact on the success of the team as a whole. Therefore, past interactions among members may have a greater influence in how individual team members rate each other than was found with ratings of leader ability. The team would then be more aware of the impact their relationships have on each other and be able to address these issues accordingly.

Data analysis programs. The present study used a combination of SAS and AMOS to estimate SRM components as well as the self-enhancement model. However, these programs are relatively costly and many organizations may not already have access to them. Previous studies have set established a tradition of using mixed-modeling in SAS to calculate SRM components. This program also has a sophisticated way to handle missing data, which was necessary for the

restructured data used in the present study. However, a new program in the open-source R community called *TripleR* (Schmukle, Schonbrodt, & Back, 2010) is also able to calculate SRM components, which makes this type of analysis readily available to organizations. The package provides helpful information on how to properly structure round-robin data and allows for missing data. This program provides estimates for all components of the SRM including *perceiver, target,* and *relationship* variance in addition to perceiver-target covariance, reliability estimates, and self-ratings accounting for both *perceiver* and *target* variance separately.

Conclusions

The seemingly ubiquitous presence that MSF assessment ratings have in modern day organizations, and the powerful role they play in shaping our understanding and support of leadership, provided the inspiration for the present study. A multiple methods approach was employed to test whether conventional use of MSF assessment data is appropriate when used in two primary ways: differentiation among leadership ability, and in the identification of development needs to guide leader self-regulation behavior. As predicted, the ways in which MSF assessment data are collected and analyzed have limited our understanding of leadership within an organization. Losing sight of the context in which leaders operate, such as changes in organizational operating processes, may change the way that leader behavior is viewed in the organization in ways that may prevent the accurate comparison of leader ratings over time. The present study also highlighted the shortcomings of using individual-level aggregate scores when a clear picture of leader ability is needed, especially for the more ambiguous socio-emotional leader behaviors. Finally, incorporating leader self-ratings into the larger Social Relations Model provided insight into several potential reasons why SOA discrepancy scores so often fail to

predict leader self-regulation activity such as goal-setting behavior, as well as leader effectiveness outcomes such as leader performance.

Overall, the results of the present study call for researchers and practitioners to rethink the modern leadership model in ways that would place more emphasis on social processes involved in leadership, and the factors that may contribute to relational effectiveness. Data used in the present study were collected from a relatively traditional corporate environment, but ever growing diversity in the workforce is likely to continue affecting the way employees and organizations think about leaders and leadership. Therefore, leadership researchers will increasingly need to test the limits of the more readily available and accepted methods of leader assessment and leader development. Results from the current study suggest that a more balanced approach is currently needed to promote leader self regulation, one in which leader ability as well as social-relational processes that support overall leadership are considered equally in assessment and development activities.

References

Allport, G. W. (1937). *Personality: A psychological interpretation.* New York: Holt.

Amundsen, S., & Martinsen, O. L. (2014). Self-other agreement in empowering leadership: Relationships with leader effectiveness and subordinates' job satisfaction and turnover intention. *The Leadership Quarterly, 25,* 784-800. doi: 10.1016/j.leaqua.2014.04.007

Antonioni, D. (1996). Designing an effective 360-degree appraisal feedback process. *Organizational Dynamics, 25 (2),* 24-38.

Ashford, S. J., & Tsui, A. S. (1991). Self-regulation for managerial effectiveness: The role of active feedback seeking. *Academy of Management Journal, 34(2),* 251-280.

Atkins, P. W. B, & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55(4),* 871-904.

Atwater, L. E., Waldman, D. A., & Brett, J. F. (2002). Understanding and optimizing muti-source feedback. *Human Resource Management, 41 (2),* 193-208. doi: 10.1002/hrm.10031

Atwater, L. E., & Yammarino, F. J. (1997). Self-other rating agreement: A review and model. *Research in Personnel and Human Resources Management, 15,* 121-174.

Atwater, L. E., & Yammarino, F. J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Personnel Psychology, 45,* 141-164.

Avolio, B. J., Sosik, J. J., Jung, D. I., & Berson, Y. (2013). Leadership models, methods, and applications: Progress and remaining blind spots. *In: Handbook of Industrial and*

Organizational Psychology, Volume 12 (Ed., S. Zedeck*), 277-307.* American Psychological Association.

Avolio, B. J., Sosik, J. J., Jung, D. I., & Berson, Y. (2003). Leadership models, methods, and applications. *In: Handbook of Industrial and Organizational Psychology,* (Ed., S. Zedeck), *277-307.* American Psychological Association.

Avolio, B. J, & Gardner, W. L. (2005). Authentic leadership development: Getting to the root of positive forms of leadership. *Leadership Quarterly, 16,* 315-338.

Barbuto, Jr., J. E., Wilmot, M. P., & Story, J. S. (2011). Self-other rating agreement and leader-member exchange (LMX). *Perceptual Motor Skills, 113(3),* 875-880. doi: 10.2466/01.03.21.PMS.113.6.875-880

Bass, B. M., & Yammarino, F. J. (1991). Congruence of self and others' leadership ratings of naval officers for understanding successful performance. *Applied Psychology: An International Review, 40 (4),* 437-454.

Bass, B. M. (1985). *Leadership and performance beyond expectations.* New York: Free Press.

Bazigos, M. (1999). *The relationship of upward feedback disparities to leader performance: Understanding "overestimation* (Ph.D.). Columbia University.

Bettin, P. J., & Kennedy, J. K. (1990). Leadership experience and leader performance: Some empirical support at last. *The Leadership Quarterly, 1 (4),* 219-228.

Bock, R.D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Management, 25 (4),* 275-285.

Boies, K., & Howell, J. M. (2006). Leader-member exchange in teams: An examination of the interaction between relationship differentiation and mean LMX in

explaining team-level outcomes. *The Leadership Quarterly, 17*, 246-257. doi: 10.1016/j.leaqua.2006.02.004

Bracken, D. W. (1997). Maximizing the uses of multi-rater feedback. In: D. Bracken, M. Dalton, R., Jako, C. McCauley, V. Pollman, *Should 360-degree Feedback be used Only for Development Purposes?* Center for Creating Leadership: Greensboro, NC.

Bradford, L. P. (1976). The laboratory method: A historical perspective. *Group & Organization Studies, 1(4),* 415-429.

Brutus, S., & Derayah, M. (2002). Multisource assessment programs in organizations: An insider's perspective. *Human Resource Development Quarterly, 13(2),* 187-202.

Brutus, S., Fleenor, J. W., & Tisak, J. (1999). Exploring the link between rating congruence and managerial effectiveness. *Canadian Journal of Administrative Sciences, 16(4),* 308-322.

Burke, C. S., Stagl, K. C., Klein, C., Goodwin, G. F., Salas, E., & Halpin, S. M. (2006). What type of leadership behaviors are functional in teams? A meta-anaysis. *Leadership Quarterly, 17(3),* 288-307.

Campion, M. A., & Lord, R. G. (1982). A control systems conceptualization of the goal-setting and changing process. *Organizational Behavior and Human Performance, X,* 265-287.

Carless, S. A., Mann, L., & Wearing, A. J. (1998). Leadership, managerial performance and 360-degree feedback. *Applied Psychology: An International Review, 47 (4),* 481-496.

Carmeli, A., Ben-Hador, B., Waldman, D. A., & Rupp, D. E. (2009). How leaders cultivate social capital and nurture employee vigor: Implications for job performance. *Journal of Applied Psychology, 94 (6),* 1553-1561.

Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for

personality-social, clinical, and health psychology. *Psychological Bulletin, 92 (1),* 111-135.

Cascio, W. F., & Valenzi, E. R. (1977). Behaviorally anchored rating scales: Effects of education and job experience of raters and rates. *Journal of Applied Psychology, 62,* 278-282.

Chalmers, R. Philip. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48 (6),* 1-29.

Chun, J. U., Yammarino, F. J., Dionne, S. D., Sosik, J. S., & Moon, H. K. (2009). Leadership across hierarchical levels: Multiple levels of management and multiple levels of analysis. *Leadership Quarterly, 20,* 689-707. doi: 10.1016/j.leaqua.2009.06.003

Church, A. H., Waclawski, J. & Burke, W. W. (2001). Multisource feedback for organization development and change. In D. W. Bracken, C. W. Timmreck, & A. H. Church (Eds.), *Handbook of Multisource Feedback* (301-317). San Francisco, CA: Jossey-Bass.

Church, A. H. (1997). Do you see what I see? An exploration of congruence in ratings from multiple perspectives. *Journal of Applied Social Psychology, 27 (11),* 983-1020.

Church, A. H. (1994). Managerial self-awareness in high performing individuals in organizations. *Dissertation Abstracts International,* 55-05B, 2028. (University Microfilms No. AAI9427924)

Collins, W. C., Raju, N. S., Edwards, J. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology, 85,* 451-461.

Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology, 68 (6),* 1152-1162.

Conway, J.M., & Huffcutt, A. I. (1997). Psychometric properties of multisource

performance ratings: A meta-analysis of subordinate, supervisor, peer and self-ratings. *Human Performance, 10,* 331-360.

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),* 44-60.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cullen, K. L., Gentry, W. A., & Yammarino, F. J. (2015). Biased self-perception tendencies: Self-enhancement/self-diminishment and leader derailment in individualistic and collectivistic cultures. *Applied Psychology: An International Review, 64 (1),* 161-207. doi: 10.1111/apps.12026

Dalton, M. A. (1997). When the purpose of using multi-rater feedback is behavior change. In W. W. Tornow & M. London (Eds.), *Should 360-degree Feedback Be Used Only for Developmental Purposes?* Greensboro, NC: Center for Creative Leadership.

Dansereau, F., Cho, J., & Yammarino, F. J. (2006). Avoiding the "Fallacy of the wrong level": A within and between analysis (WABA) approach. *Organizational Management, 31 (5),* 536-577. doi: 10.1177/1059601106291131

Day, D. V., Fleenor, J. W., Atwater, L. E., Sturm, R. E., & McKee, R. A. (2014). Advances in leader and leadership development: A review of 25 years of research and theory. *The Leadership Quarterly, 25,* 63-82. doi: 10.1016/j.leagua.2013.11.004

Day, D. V. (2000). Leadership development: A review in context. *Leadership Quarterly, 11(4),*

581-613.

Dionne, S. D., Gupta, A., Sotak, K. L., Shirreffs, K. A., Serban, A., Hao, C., Kim, D. H., & Yammarino, F. J. (2014). A 25-year perspective on levels of analysis in leadership research. *The Leadership Quarterly, 25,* 6-35. doi: 10.1016/j.leaqua.2013.11.002

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after likert: Thurstone was right! *Industrial and Organizational Psychology, 3,* 465-476.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (2nd ed., Vol.1, pp. 577-636). Palo Alto, CA: Consulting Psychologists Press.

Edwards, J. R. (1992). A cybernetic theory of stress, coping and well-being in organizations. *Academy of Management Review, 17(2),* 238-274.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Emerson, R. M. (1976). Social exchange theory. Annual Review of Sociology, 335-362.

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

Farh, J., & Dobbins, G. (1989). Effects of comparative performance information on the accuracy of self-ratings and agreement between self and supervisor ratings. *Journal of Applied Psychology, 74,* 606-610.

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

Fleenor, J. W., Smither, J. W., Atwater, L. E., Braddy, P. W., & Sturm, R. E. (2010). Self-other rating agreement in leadership: A review. *The Leadership Quarterly, 21,* 1005-1034. doi: 10.1016/j.leaqua.2010.10.006

Fleenor, J. W., McCauley, C. D., & Brutus, S. (1996). Self-other rating agreement and leader effectiveness. *Leadership Quarterly, 7*, 487-506.

Fletcher, C., & Baldry, C. (2000). A study of individual differences and self-awareness in the context of multi-source feedback. *Journal of Occupational and Organizational Psychology, 73*, 303-319.

Froming, W., & Carver, C. (1981). Divergent influences of private and public self-consciousness in a compliance paradigm. *Journal of Research in Personality, 15*, 159-171.

Goffin, R. D., & Anderson, D. W. (2007). The self-rater's personality and self-other disagreement in multi-source performance ratings: Is disagreement healthy? *Journal of Managerial Psychology, 22*, 271-289.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20(4)*, 369-377.

Gooty, J., & Yammarino, F. J. (2011). Dyads in organizational research: Cenceptual issues and multilevel analyses. *Organizational Research Methods, 14(3)*, 456-483. doi: 10.1177/1094428109358271

Graen, G. B., Hui, C., Taylor, E. A. (2006). Experience-based learning about LMX leadership and fairness in project teams: A dyadic directional approach. *Academy of Management 5 (4)*, 448-460.

Graen, G. B., & Uhl-Bien, M. (1995). Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory over the last 25 years: Applying a multi-level multi-domain approach. *Leadership Quarterly, 6*, 219-247.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist, 35 (7)*, 603-618.

Greguras, G. J., Robie, C., Born, M. P., & Koenigs, R. J. (2007). A social relations analysis of team performance ratings. *International Journal of Selection & Assessment, 15,* 434-448. doi:10.1111/j.1468-2389.2007.00402.x

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.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* SAGE Publications: London, UK.

Harrington, A. (2015). *Numbers, words and anonymity in 360-degree feedback: A qualitative study* (Ph.D.). Loughborough University Institutional Repository.

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

Hennen, M.E., & Barnes-Farrell, J. L. (1997). *Appraisal in self-managing work groups: a social relations model".* Presented at the 12th Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO, April.

Highhouse, S. (2002). Assessing the candidate as a whole: A historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology, 55,* 363-396.

Hiller, N. J., DeChurch, L.A., Murase, T., & Doty, D. (2011). Searching for outcomes of leadership: A 25-year review. *Journal of Management, 37 (4),* 1137-1177. doi: 10.1177/0149206310393520

Hollenbeck, G. P, & Hall, D. T. (2004). Self-confidence and leader performance. *Organizational*

Dynamics, 33(3), 254-269. doi: 10.1016/j.orgdyn.2004.06.003

Johnston, W. A. (1967). Individual performance and self-evaluation in a simulated team. *Organizational Behavior and Human Performance, 2,* 309-328.

Kanfer, R. (1990). Motivation and individual differences in learning: An integration of developmental, differential and cognitive perspectives. *Learning and Individual Differences, 2,* 221-239.

Kavanagh, M. J., MacKinny, A., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin, 75,* 34-39.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis.* New York: Guilford Press.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic Data Analysis,* Guilford Press: New York, NY.

Kenny, D. A., & Livi, S. (2009). A componential analysis of leadership using the social relations model. In: *Multi-Level Issues in Organizational Behavior and Leadership, Vol. 8.* (pp. 147-191). Cambridge, MA: Emerald Publishing Group.

Kenny, D. A., & West, T. V. (2010). Similarity and agreement in self- and other perception: A meta-analysis. *Personality and Social Psychology Review, 14(2),* 196-213.

Kenny, D. A., & Zaccaro, S. J. (1983). An estimate of variance due to traits in leadership. *Journal of Applied Psychology, 68 (4),* 678-685.

Kempster, S., & Parry, K. (2014). Exploring observational learning in leadership development for managers. *Journal of Management Development, 33 (3),* 164-181. doi: 10.1108/JMD-01-2012-0016

Kornhauser, A. W. (1923). A statistical study of a group of specialized office workers.

Journal of Personnel Research, 2, 103-123.

Kulas, J. T., & Finkelstein, L. M. (2007). Content and reliability of discrepancy-defined self-awareness in multisource feedback. *Organizational Research Methods, 10 (3),* 502-522.

Kwan, V. S. Y., John, O. P., Kenny, D. A., Hond, M. H., & Robins, R. W. (2004). Reconceptualizing individual differences in self-enhancement: An interpersonal approach. *Psychological Review, 111 (1),* 94-110. doi:10.1037/0033-295X.111.1.94

LeDoux, J. A., Gorman, C. A., & Woehr, D. J. (2012). The impact of interpersonal perceptions on team processes: A social relations analysis. *Small Group Research, 43 (3),* 356-382. doi:10.1 177/1046496411425190

Link, H. C. (1920). The applications of psychology to industry. *Applications of Psychology to Industry,* 335- 336.

London, M. (2001). The great debate: Should multisource feedback be used for administration or development only? In D. W. Bracken, C. W. Timmreck, & A. H. Church (Eds.), *Handbook of Multisource Feedback* (15-32). San Francisco, CA: Jossey-Bass.

London, M., & Beatty, R. W. (1993). 360-degree feedback as a competitive advantage. *Human Resource Management, 32,* 352-373.

London, M., & Tornow, W. W. (1998). 360-degree feedback: More than a tool! In D.W. Bracken and others (Eds.), *Should 360-Degree Feedback Be Used Only for Developmental Purposes?* Greensboro, NC: Center for Creative Leadership.

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., De Vader, C. L., & Alliger, G. M. (1986). A meta-analysis of the relation between personality traits and leadership perceptions: An application of validity generalization procedures. *Journal of Applied Psychology, 71 (3),* 402-410.

Ma. L., & Qu, Q. (2010). Differentiation in leader-member exchange: A hierarchical linear modeling approach. *The Leadership Quarterly, 21,* 733-744. doi: 10.1016/j.leaqua.2010.07.004

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.

Malloy, T. E., & Kenny, D. A. (1986). The social relations model: An integrative method for personality research. *Journal of Personality, 54 (1),* 199-225.

Manz, C. C. (1986). Self-leadership: Toward an expanded theory of self-influence processes in organizations. *Academy of Management Review, 11(3),* 585-600.

Markham, S. E., Smith, J. W., Markham, I. S., & Braekkan, K. F. (2014). A new approach to analyzing the Achilles' heel of multisource feedback programs: Can we really trust ratings of leaders at the group level of analysis? *Leadership Quarterly, 25,* 1120-1142. doi: 10.1016/j.leaqua.2014.10.003

Maylett, T. (2009). 360-degree feedback revisited: The transition from development to appraisal. *Compensation and Benefits Review, 41 (5),* 52-59.

Morgan, A., Cannan, K., & Cullinane, J. (2005). 360-degree feedback: A critical enquiry. *Personnel Review, 34 (6),* 663-680.

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, 557-576.

Mueller, B. H., & Lee, J. (2002). Leader-member exchange and organizational communication satisfaction in multiple contexts. *The Journal of Business Communication, 39 (2),* 220-244.

Mumford, T. V., Campion, M. A., & Morgeson, F. P. (2007). Leadership skills strataplex: Leadership skill requirements across organizational levels. *Leadership Quarterly, 18(2),* 154-166. doi: 10.1016/j.leaqua.2007.01.005

Mumford, M., Zaccaro, S. J., Johnson, J. F., Diana, M., Gilbert, J.A., & Threlfall, K. (2000). Patterns of leader characteristics: Implications for performance and development. *The Leadership Quarterly, 11(1),* 115-133.

Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology, 1,* 148-160.

Murphy, K., & Cleveland, J. (1995). *Understanding Performance Appraisal: Social, Organizational, and Goal-Oriented Perspectives.* Thousand Oaks, CA: Sage.

Penny, J. (2010). Differential item functioning in an international 360-degree assessment: Evidence of gender stereotype, environmental complexity, and organizational contingency. *European Journal of Work and Organizational Psychology, 10(3),* 245-271. doi: 10.1080/13594320143000663

Penny, J. (2003). Exploring differential item functioning in a 360-degree assessment: Rater source and method of delivery. *Organizational Research Methods, 6(1),* 61-79.

Porath, C. L., & Bateman, T. S. (2006). Self-regulation: From goal orientation to job performance. *Journal of Applied Psychology, 91 (1),* 185-192. doi: 10.1037/0021-91010.91.1.185

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review, 11(3),* 781-799. doi: 10.1073/0033-259X.111.3.781

Raju, N. S., & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.). *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 156-188). San Francisco, CA: Jossey-Bass.

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17 (5),* 1-25. URL: http://www.jstatsoft.org/v17/i05

Rowold, J., Borgmann, L., & Diebig, M. (2014). A "Tower of Babel"? – Interrelations and structure of leadership constructs. *Leadership & Organization Development Journal, 36 (2),* 137-160. doi: 10.1108/LODJ-01-2013-0009

Samajima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement.*

Schmukle, S. C., Schönbrodt, F. D., & Back, M. D. (2010). TripleR: A package for round robin analyses using R (version 1.1.0). Retrieved from http://www.persoc.net/ToolBox/TripleR.

Scullen, S. E. (1997). When ratings from one source have been averaged, but ratings from another source have not: Problems and solutions. *Journal of Applied Psychology, 82,* 880-888.

Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for though, *Perspectives*

on Psychological Science, 3 (2), 102-116.

Shipman, A. S., & Mumford, M. D. (2011). When confidence is detrimental: Influence of overconfidence on leadership effectiveness. *Leadership Quarterly, 22,* 649-665.

Shore, T. H., Adams, J. S., & Tashchian, A. (1998). Effects of self-appraisal information, appraisal purpose, and feedback target on performance appraisal ratings. *Journal of Business and Psychology, 12,* 283-298.

Sin, H. P., Nahrgang, J. D., & Moregeson, F. P. (2009). Understanding why they don't see eye-to-eye: An examination of leader-member exchange (LMX) agreement. *Journal of Applied Psychology. 94 (4),* 1048-1057.

Slater, R., & Coyle, A. (2014). The governing of the self/the self-governing self: Multi-rater/source feedback and practices 1940-2011. *Theory & Psychology, 24 (2),* 233-255.

Smith, T. J. (2015). Variability in human work performance: Interaction with complex sociotechnical systems. In: Smith, T.J., Henning, R. A., Wade, M. G., & Fisher, T. *Variability in Human Performance,* 211-283. Boca Raton, FL: Taylor & Francis Group.

Smith, T.J., & Smith, K.U. (1987). Feedback-control mechanisms of human behavior. In G. Salvendy (Ed.), *Handbook of human factors* (pp. 251-293). New York: Wiley.

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,* 33-66.

Sosik, J. J., Potosky, D., & Jung, D. I. (2002). Adaptive self-regulation: Meeting others' expectations of leadership and performance. *Journal of Social Psychology, 142(2),* 211-232. doi: 10.1080/00224540209603896

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions

about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91(1),* 25-39. doi: 10.1037/0021-9010.91.1.25

Stogdill, R. (1974). *Handbook of leadership* (1st Edition). New York: Free Press.

Strang, S. E., & Kuhnert, K. W. (2009). Personality and leadership developmental levels as predictors of leader performance. *The Leadership Quarterly, 19(3),* 360-371.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103 (2),* 193-210.

Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin, 116 (1),* 21-27.

Taylor-Bianco, A., & Schermerhorn, Jr., J. (2000). Self-regulation, strategic leadership and paradox in organizational change. *Journal of Organizational Change Management, 19 (4),* 457-470. doi: 10.1108/09534810610676662

Tornow, W. W. (1998). Forces that affect the 360-degree feedback process. In: W. Tornow, & M. London, *Maximizing the Value of 360-Degree Feedback: A Process for Successful Individual and Organizational Development.* Center for Creative Leadership: Greensboro, NC.

Tse, H. M., Dasborough, M., & Ashkanasay, N. M. (2008). A multilevel analysis of team climate and interpersonal exchange relationships at work. *Leadership Quarterly, 19,* 195-211.

van der Kam, N. A., Janssen, O., van der Vegt, G. S., & Stoker, J. I. (2013). The role of vertical conflict in the relationship between leader self-enhancement and leader performance. *The Leadership Quarterly, 25,* 267-281. doi: 10.1016/j.leaqua.2013.08.007

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557-574.

Wheelan, S. A., & Johnston, F. (1996). The role of informal member leaders in a system containing formal leaders. *Small Group Research, 27 (1),* 33-55.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33 (42),* 43-57. doi: 10.1177/0146621607314044

Yammarino, F. J., & Dansereau, F. (2002). Individualized leadership. *Journal of Leadership and Organizational Studies, 9 (1),* 90-99.

Yammarino, F. J. (2003). Modern data analytic techniques for multisource feedback. *Organizational Research Methods, 6 (1),* 6-14.

Tables

Table 1

Proportion of Variance as Measured by EFA Explained by First Factor

	Task-related	Socio-emotional
Middle Tier	0.57	0.59
Foundational Tier	0.62	0.60
Top Tier	0.52	0.51

Table 2

Exploratory Factor Analysis: 2-Factor Solution with Oblique Rotation

	Factor 1	Factor 2
Task-related Items		
q1	**0.86**	
q2	**0.8**	
q3	**0.71**	
q4	**0.76**	
q5	**0.66**	
q6	**0.8**	
q7	**0.78**	
q8	**0.81**	
q9	**0.65**	
q10	**0.78**	
q11	**0.72**	
q12	**0.53**	
q13	**0.64**	
q14	**0.68**	
q15	**0.77**	
q16	**0.76**	
q17	**0.67**	
q18	**0.6**	
q19	**0.74**	
q20	**0.77**	
Socio-emotional Items		
q21		**0.87**
q22		**0.84**
q23		**0.88**
q24		**0.8**
q25		**0.77**
q26		**0.68**
q27		**0.8**
q28	**0.56**	
q29	**0.43**	
q30	**0.67**	
q31	**0.65**	
q32	**0.53**	
q33	**0.52**	
q34	**0.54**	
q35	**0.59**	
q36		**0.43**
q37	**0.5**	
q38		**0.54**
q39	**0.54**	
q40	**0.41**	
q41	**0.49**	
q42	**0.42**	**0.41**
q43	**0.51**	
q44		**0.52**
q45	**0.47**	
q46		**0.47**
q47	**0.54**	

Note: Factor loadings < .30 were not reported

Table 3

Two-Factor Item Fit

CFI	TLI	RMSEA
0.996	0.990	0.068

Table 4

Fit of Graded Response Model

	Task-related			Socio-emotional		
	CFI	TLI	RMSEA	CFI	TLI	RMSEA
Middle Tier						
Peer	0.91	0.90	0.07	0.93	0.90	0.07
Direct Report	0.92	0.91	0.06	0.91	0.90	0.07
Manager	0.86	0.84	0.07	0.82	0.80	0.08
Foundational Tier						
Peer	NA	NA	NA	0.82	0.80	0.09
Direct Report	NA	NA	NA	0.89	0.86	0.08
Manager	NA	NA	NA	0.72	0.67	0.11
Top Tier						
Peer	0.93	0.91	0.05	0.8	0.78	0.09
Direct Report	0.89	0.87	0.07	0.85	0.83	0.08
Manager	0.93	0.91	0.08	0.82	0.80	0.09

Table 5

Ranked Item Discrimination: *Middle Tier* Leaders

Peers				Direct Reports				Managers			
Percent Tailored Item Type	70% Task	Percent Tailored Item Type	54% Social	Percent Tailored Item Type	50% Task	Percent Tailored Item Type	54% Social	Percent Tailored Item Type	60% Task	Percent Tailored Item Type	54% Social
	2.449		3.164		2.752		2.946		2.24		2.263
Core	2.423	Core	2.949		2.728		2.768	Core	2.107		2.117
	2.419		2.759	Core	2.666		2.632	Core	2.077	Core	2.027
	2.404		2.729	Core	2.602	Core	2.628	Core	2.007		1.904
Core	2.36	Core	2.625		2.495		2.566		1.964		1.864
Core	2.343	Core	2.549	Core	2.485		2.517		1.903	Core	1.863
	2.215		2.499	Core	2.474	Core	2.458		1.86	Core	1.844
	2.208	Core	2.42	Core	2.449	Core	2.455		1.856		1.733
	2.165	Core	2.397		2.433	Core	2.42	Core	1.769	Core	1.731
	2.138		2.388		2.394		2.347		1.748	Core	1.707
Core	2.113	Core	2.366	Core	2.393		2.312	Core	1.729		1.695
Core	2.101		2.337		2.375	Core	2.294		1.719		1.694
Core	2.101		2.306		2.259	Core	2.28	Core	1.71	Core	1.687
Core	2.084	Core	2.275		2.179		2.218	Core	1.694		1.687
Core	2.081		2.168	Core	2.138		2.206		1.681		1.621
	1.999		2.165		2.138	Core	2.187	Core	1.638	Core	1.606
	1.899		2.164	Core	2.092	Core	2.154		1.455	Core	1.591
	1.894	Core	2.126		2.063	Core	2.092	Core	1.451	Core	1.538
Core	1.882	Core	2.122	Core	2.02		2.091		1.346	Core	1.513
	1.647	Core	2.087		1.784	Core	2.081		1.305		1.498
		Core	2.065			Core	1.99				1.484
			1.957			Core	1.979			Core	1.425
			1.927				1.938			Core	1.3
		Core	1.862				1.874			Core	1.242
		Core	1.862			Core	1.794				1.234
		Core	1.763			Core	1.75				1.22
	2.449		1.763				1.638			Core	1.13

Table 6

Ranked Item Discrimination: *Top Tier* Leaders

Peers				Direct Reports				Managers			
Percent Tailored Item Type	78% Task	Percent Tailored Item Type	43% Social	Percent Tailored Item Type	78% Task	Percent Tailored Item Type	43% Social	Percent Tailored Item Type	67% Task	Percent Tailored Item Type	21% Social
	2.481		2.743		2.463		2.771		16.425	Core	2.537
	2.434	Core	2.554		2.456	Core	2.622	Core	1.257		2.43
	2.408		2.46	Core	2.445	Core	2.484	Core	1.018	Core	1.987
	2.388	Core	2.372		2.411		2.48		0.902	Core	1.938
Core	2.354		2.318		2.394		2.427		0.628		1.932
	2.3	Core	2.25	Core	2.385	Core	2.4		0.611	Core	1.909
	2.293	Core	2.21	Core	2.364	Core	2.368		0.551	Core	1.861
	2.286	Core	2.181		2.324		2.343		0.528	Core	1.837
Core	2.225	Core	2.155		2.228		2.332	Core	0.521	Core	1.825
Core	2.192	Core	2.145		2.215	Core	2.29	Core	0.505	Core	1.812
Core	2.188		2.122		2.191	Core	2.282		0.503		1.787
	2.141	Core	2.051	Core	2.157	Core	2.275		0.498	Core	1.697
Core	2.064		2.039	Core	2.063	Core	2.25		0.484	Core	1.619
	1.956		2.027	Core	2.007	Core	2.219	Core	0.481		1.609
Core	1.888	Core	1.973	Core	1.984		2.06	Core	0.46		1.608
Core	1.786	Core	1.966	Core	1.923	Core	2.013		0.401		1.566
	1.78		1.961		1.914		2.006	Core	0.369		1.537
Core	1.745		1.945	Core	1.748		1.988	Core	0.252		1.466
Core	1.535		1.925		1.673		1.964	Core	0.15		1.419
		Core	1.902				1.928				1.413
			1.767			Core	1.899			Core	1.402
		Core	1.752				1.89				1.381
			1.719			Core	1.889				1.362
		Core	1.715			Core	1.885			Core	1.307
			1.7				1.799			Core	1.28
			1.587			Core	1.63				1.271
		Core	1.554				1.591				1.215

Table 7

Items with DIF from 2009- 2013

Direct Reports				Peers			
Task-related Items		Socio-emotional Items		Task-related Items		Socio-emotional Items	
Item Number	Type of DIF	Item Number	Type of DIF	Item Number	Type of DIF	Item Number	Type of DIF
6	C,NU	1	C,NU	3	C,NU	1	C,NU
7	C,NU	3	C,NU	4	C,NU	2	C,NU
11	C,NU	26	C,NU	6	C,NU	3	C,NU
				11	C,NU		
				12	C,NU		
				15	C,NU		
				16	C,NU		
				17	C,NU		
				19	C,NU		

Note: C,NU= Crossing, non-uniform

Table 8

SRM Nested Model Comparison

	Task-Related -2 Log Likelihood		Socio-emotional -2 Log Likelihood	
Model				
Random Effects	2330.9		2645.7	
Group=0				
Dyad=0	2331.4		2644.7	
χ^2 (2)	0.5	ns	χ^2 (2) 1	ns

Table 9

SRM Variance Components of Peer Ratings

	Task-related		Socio-emotional	
Random Effects	Absolute	Relative	Absolute	Relative
Perceiver	0.076	26%	0.097	27%
Target	0.136	46%	0.165	46%
Relationship	0.081	28%	0.1	28%
Fixed Covariates	Estimate	*p*	Estimate	*p*
Intercept	3.771	< .001	3.828	< .001
Target Gender	-0.027	.500	-0.01	.791
Perceiver Gender	-0.116	.006	-0.53	.254
Target*Perceiver Gender Interaction	0.213	.002	0.17	.025
Target Tenure	0.004	.082	0.003	.190
Perceiver Tenure	0.001	.689	0.001	.601

Table 10

Leader Self-enhancement and Leader Effectiveness Outcomes

	Task		Social	
	β	*p*	*β*	*p*
Self-Ratings				
k (Social Comparison)	0.169	0.067	0.404	< .001
q (Self-Insight Index)	0.587	< .001	0.536	< .001
Effectiveness Outcomes				
Leader Quality				
Perceiver	0.407	< .001	0.527	< .001
Target	0.082	0.232	0.010	0.881
Self-Enhancement	0.196	0.002	0.206	< .001
Termination				
Perceiver	1.610	< .001	1.661	< .001
Target	-.143	.573	-.134	.564
Self-Enhancement	-.257	.271	-.635	.002
Promotions				
Perceiver	.789	.001	.469	.046
Target	-.017	.908	.154	.246
Self-Enhancement	.116	.378	.015	.897
Performance				
Perceiver	.295	.049	.687	< .001
Target	-.109	.211	-.082	.267
Self-Enhancement	-.026	.745	-.082	.209

Figures

Figure 1

Part II Parallel Analysis - Peers

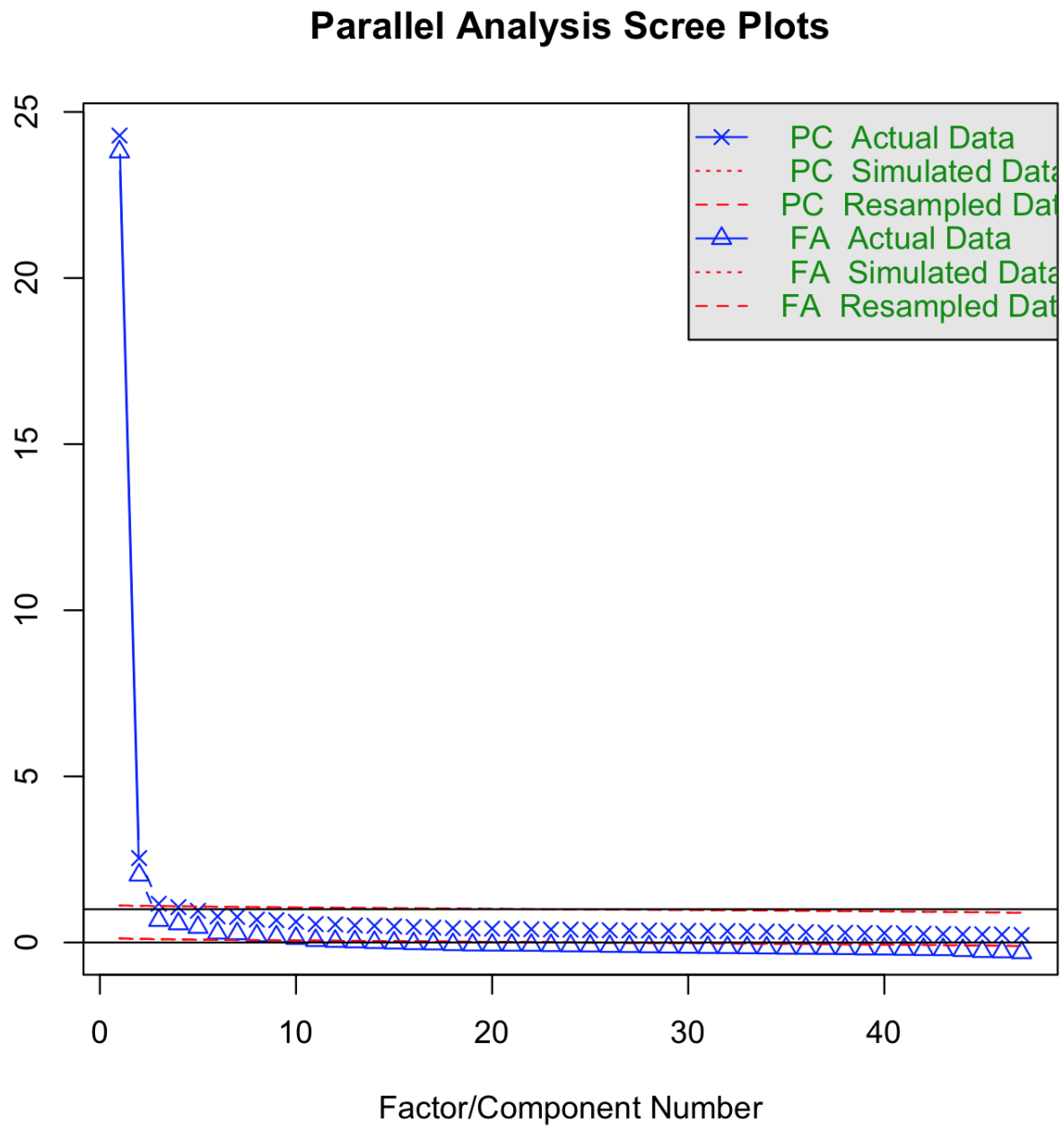


Figure 2

Part II Parallel Analysis – Direct Reports

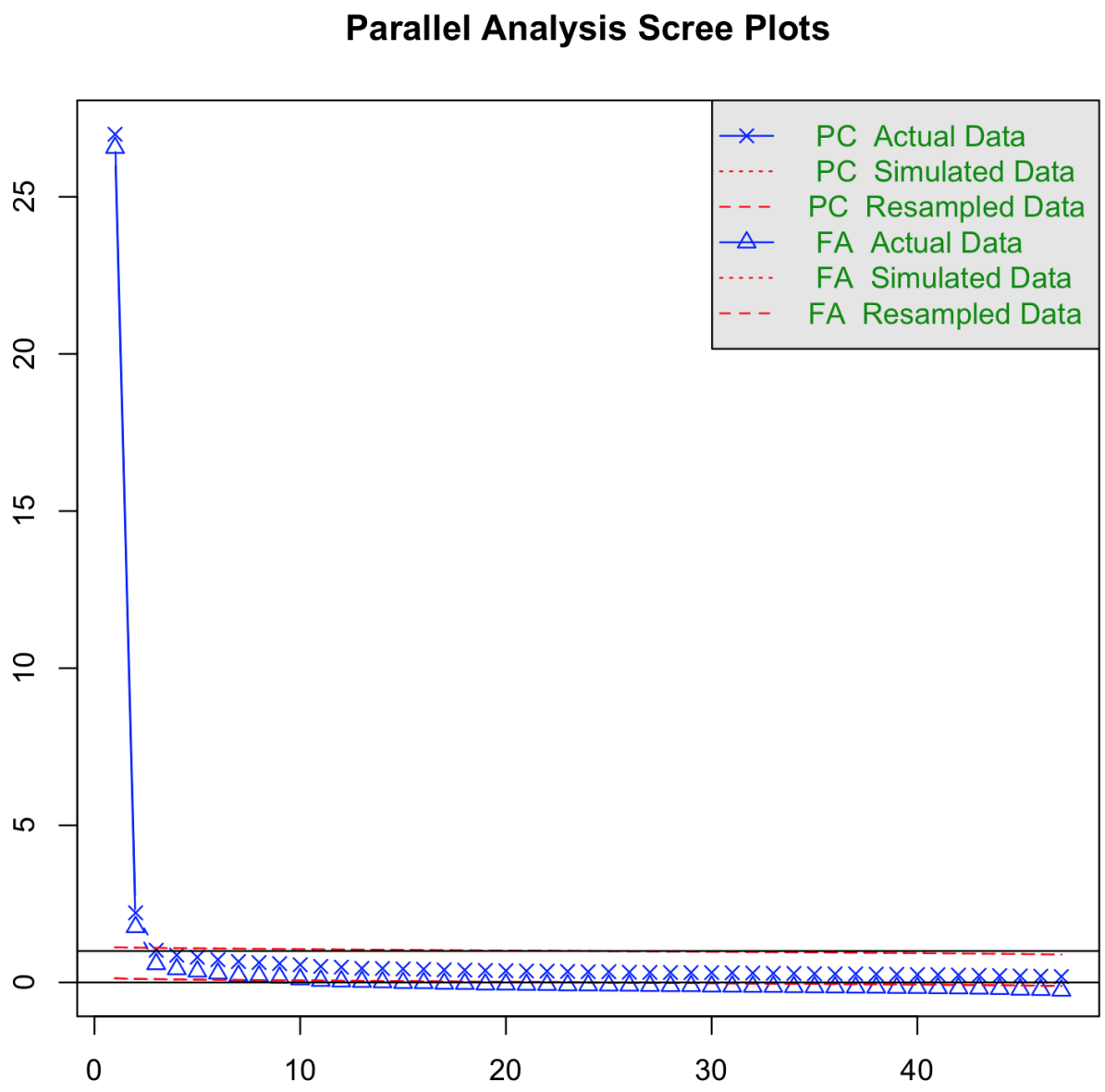


Figure 3

Peers Task-Related Test Information Curve 2009 – 2013

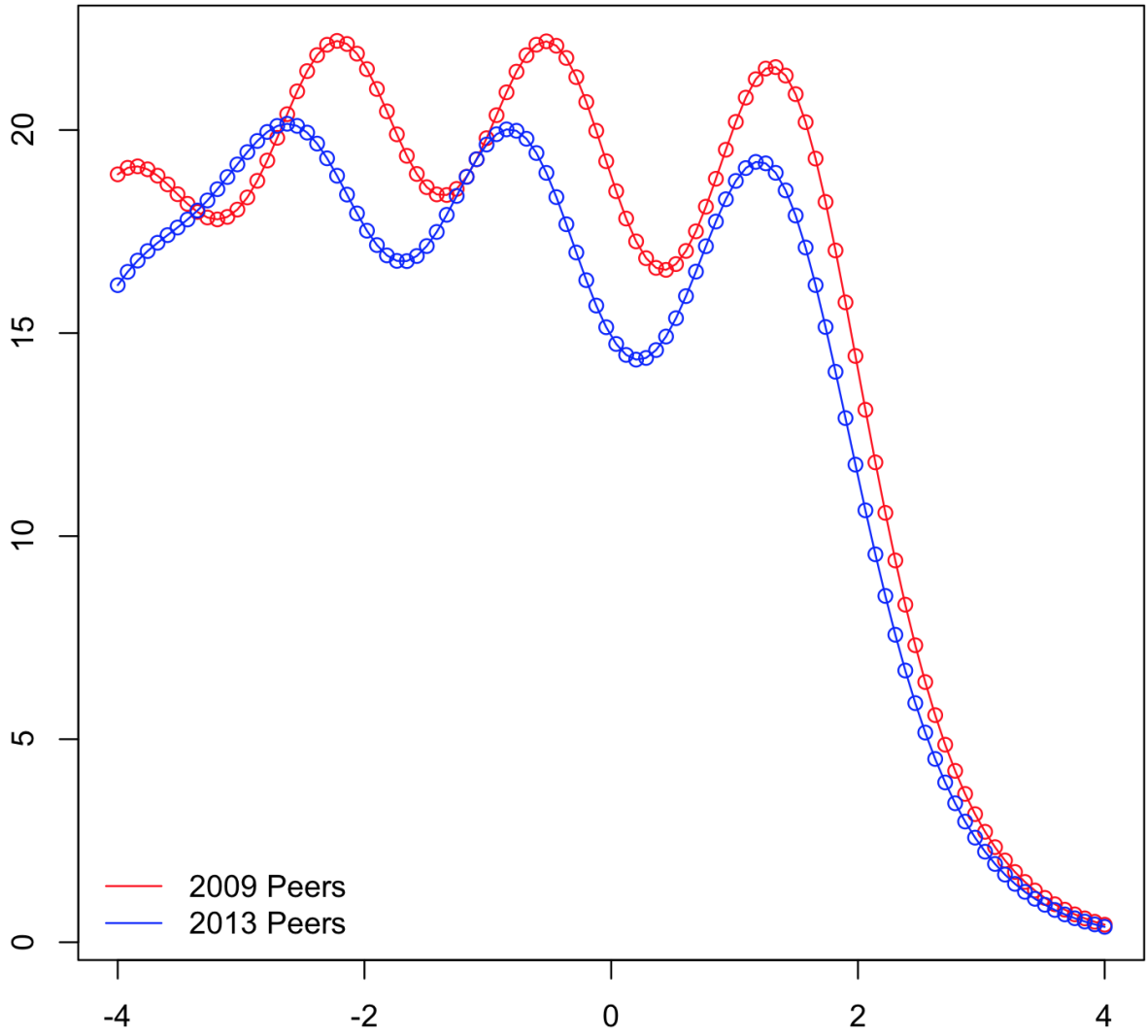


Figure 4

Direct Reports Task-Related Test Information Curve 2009 – 2013

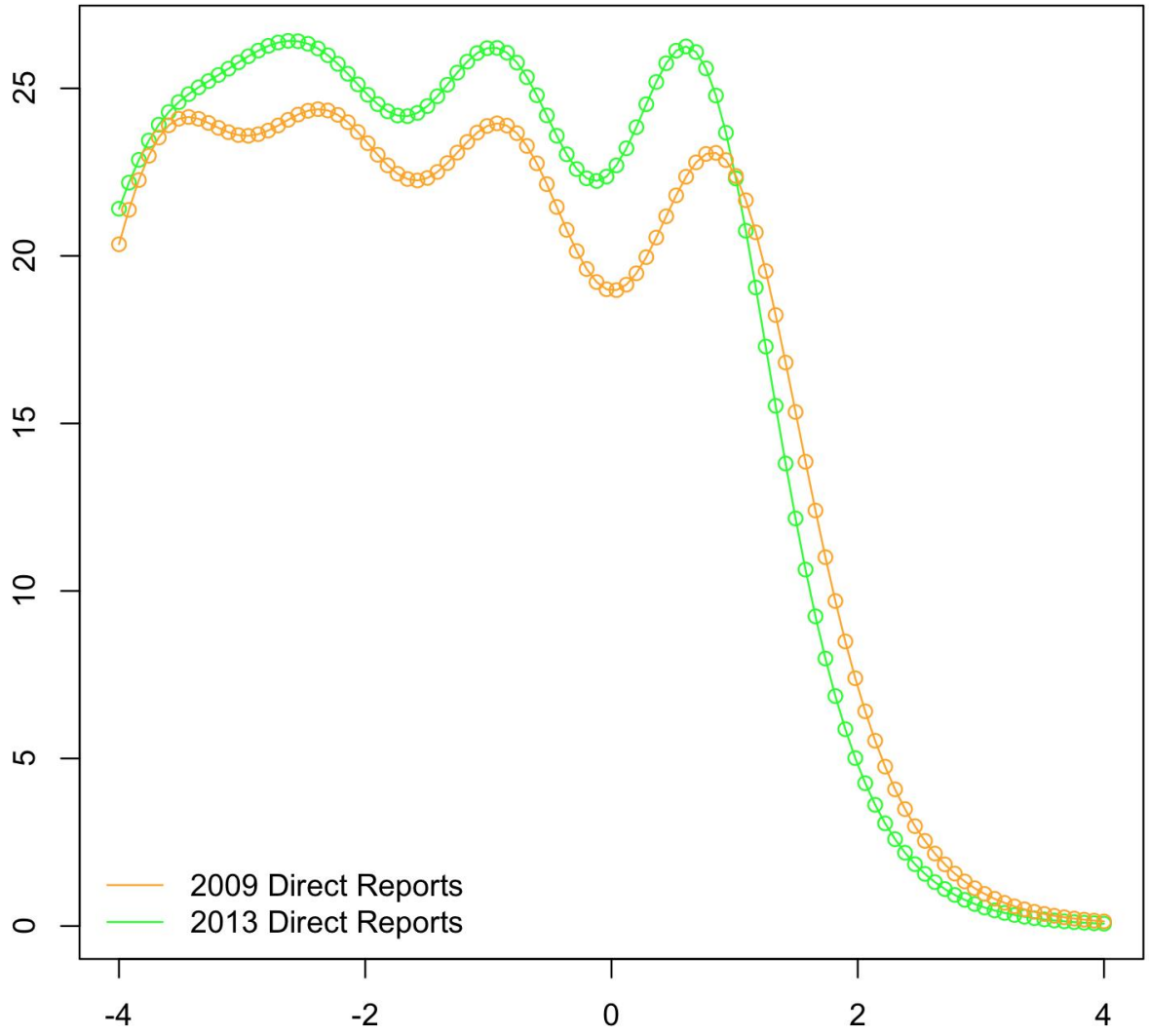


Figure 5

Peers Socio-emotional Test Information Curve 2009 – 2013

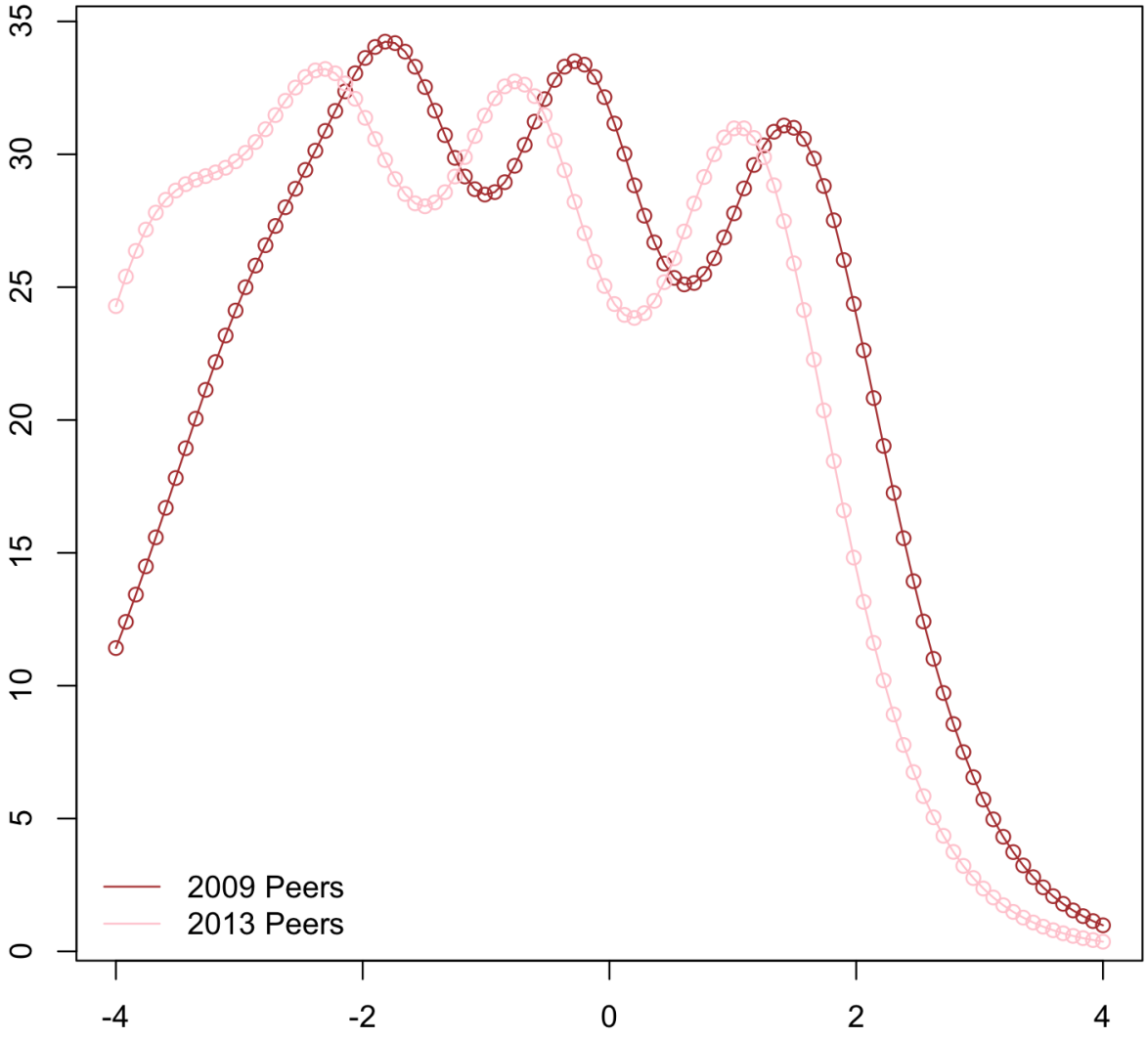


Figure 6

Direct Reports Socio-emotional Test Information Curve 2009 – 2013

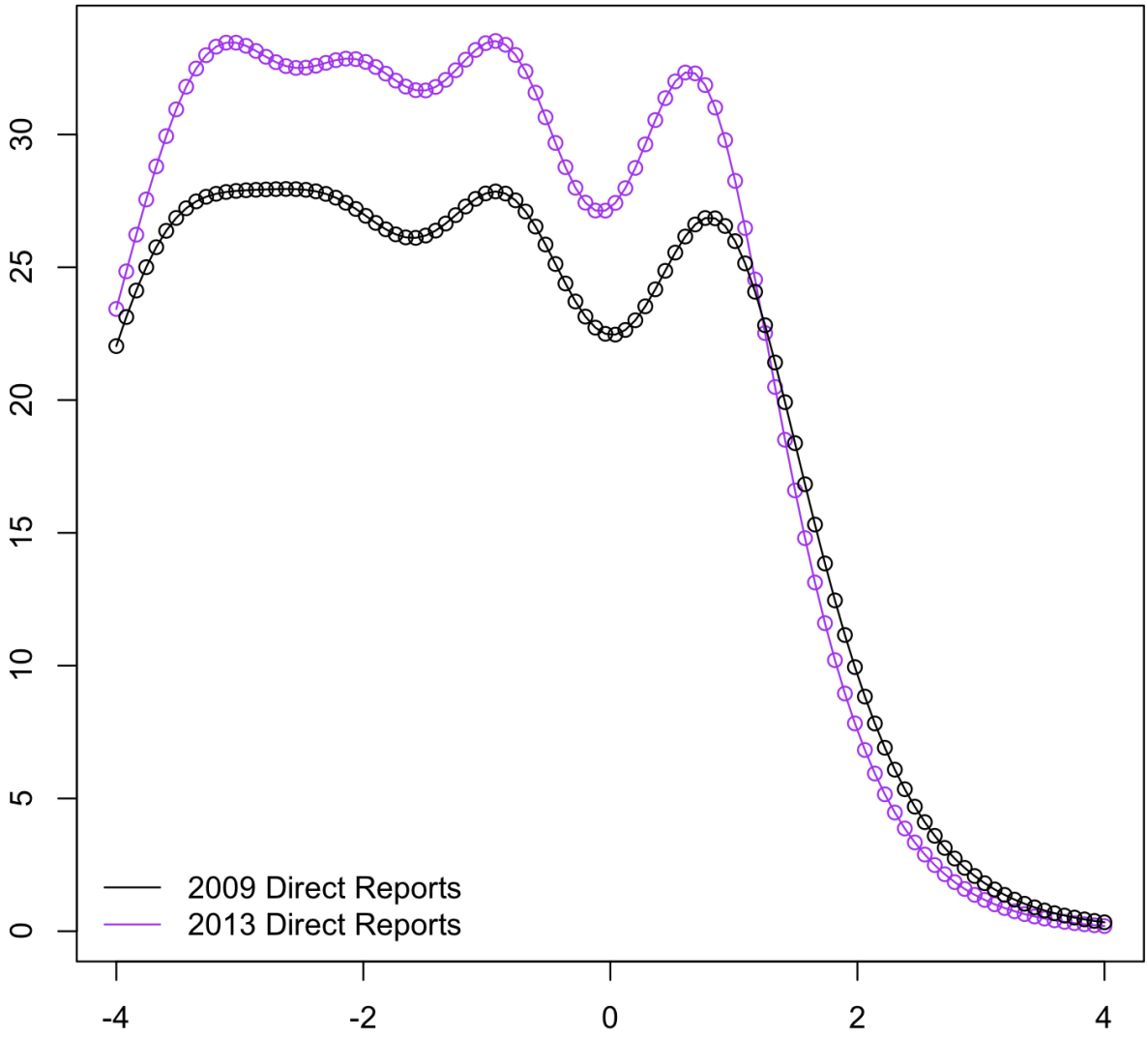
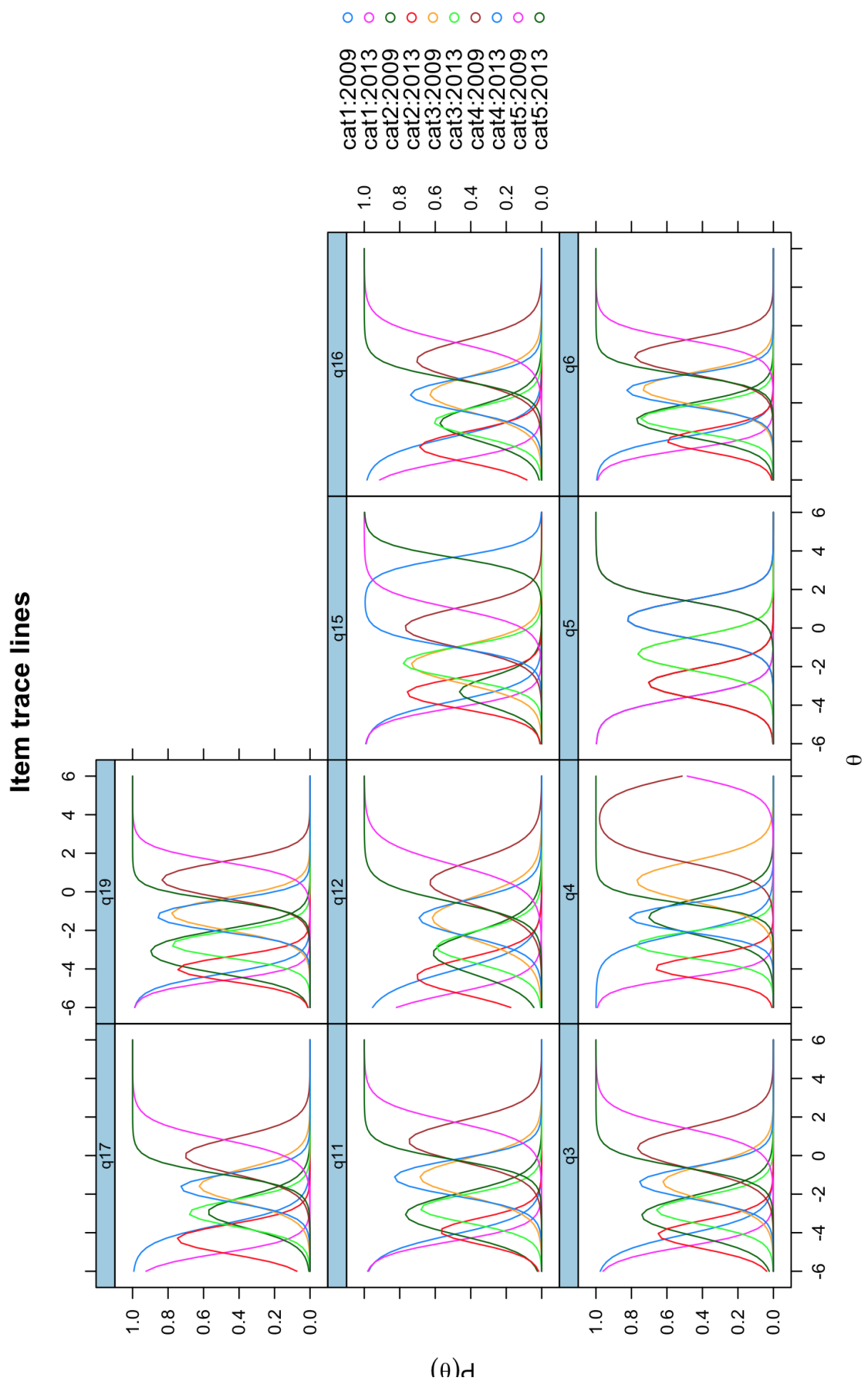


Figure 7

Task Differential Item Functioning – Peer Ratings



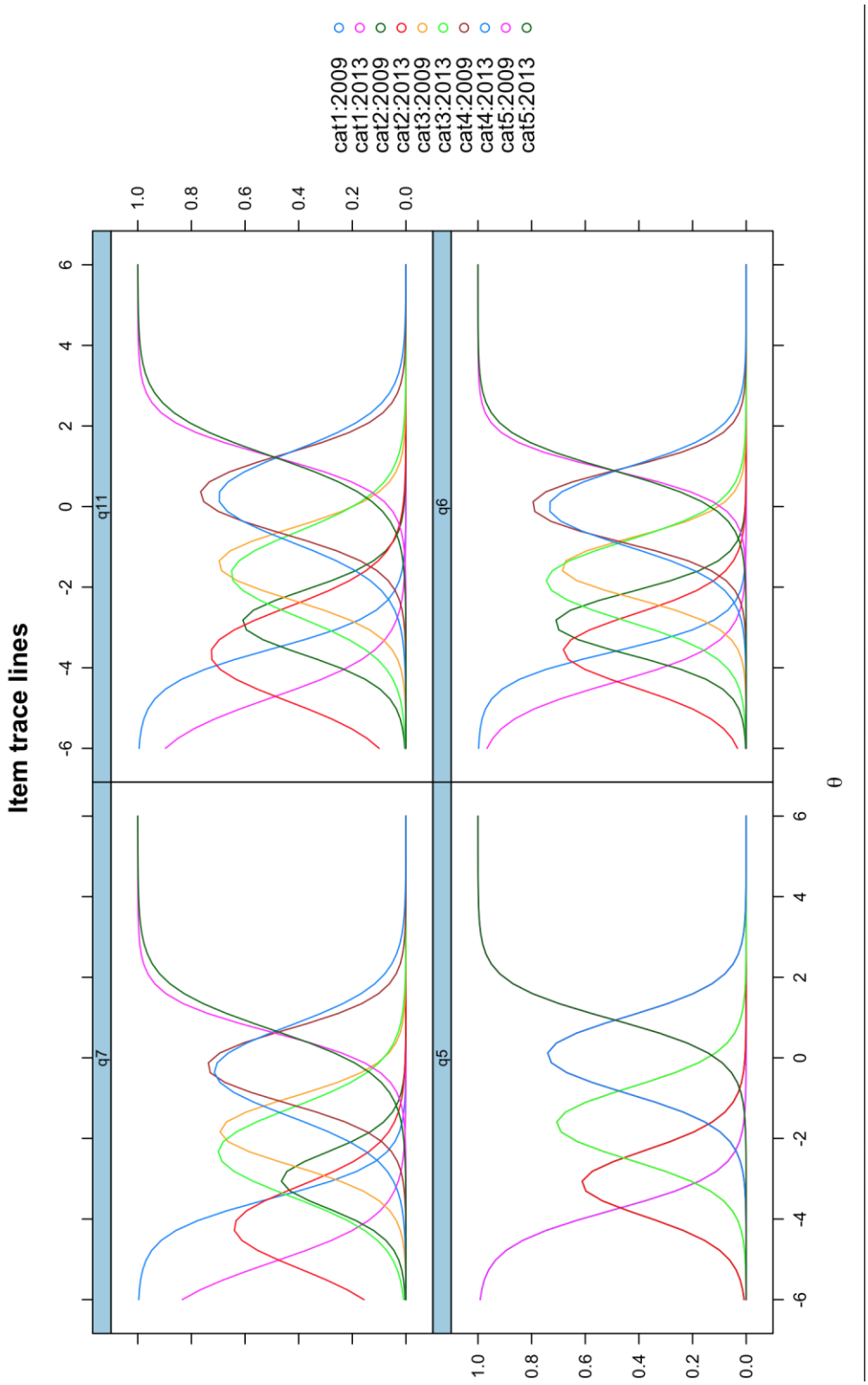


Figure 8

Task Related DIF – Direct Report Ratings

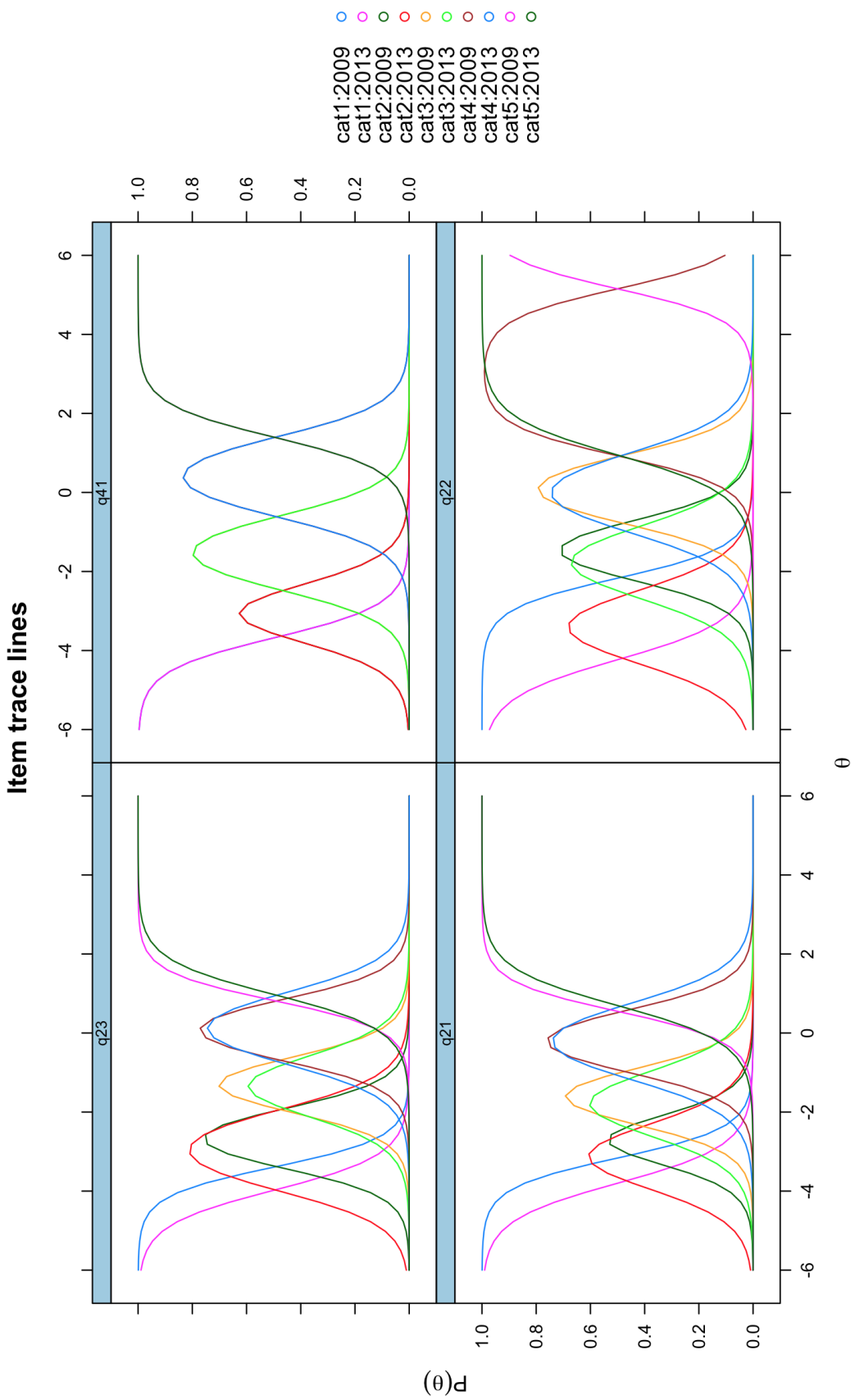


Figure 9

Socio-emotional DIF – Peer Ratings

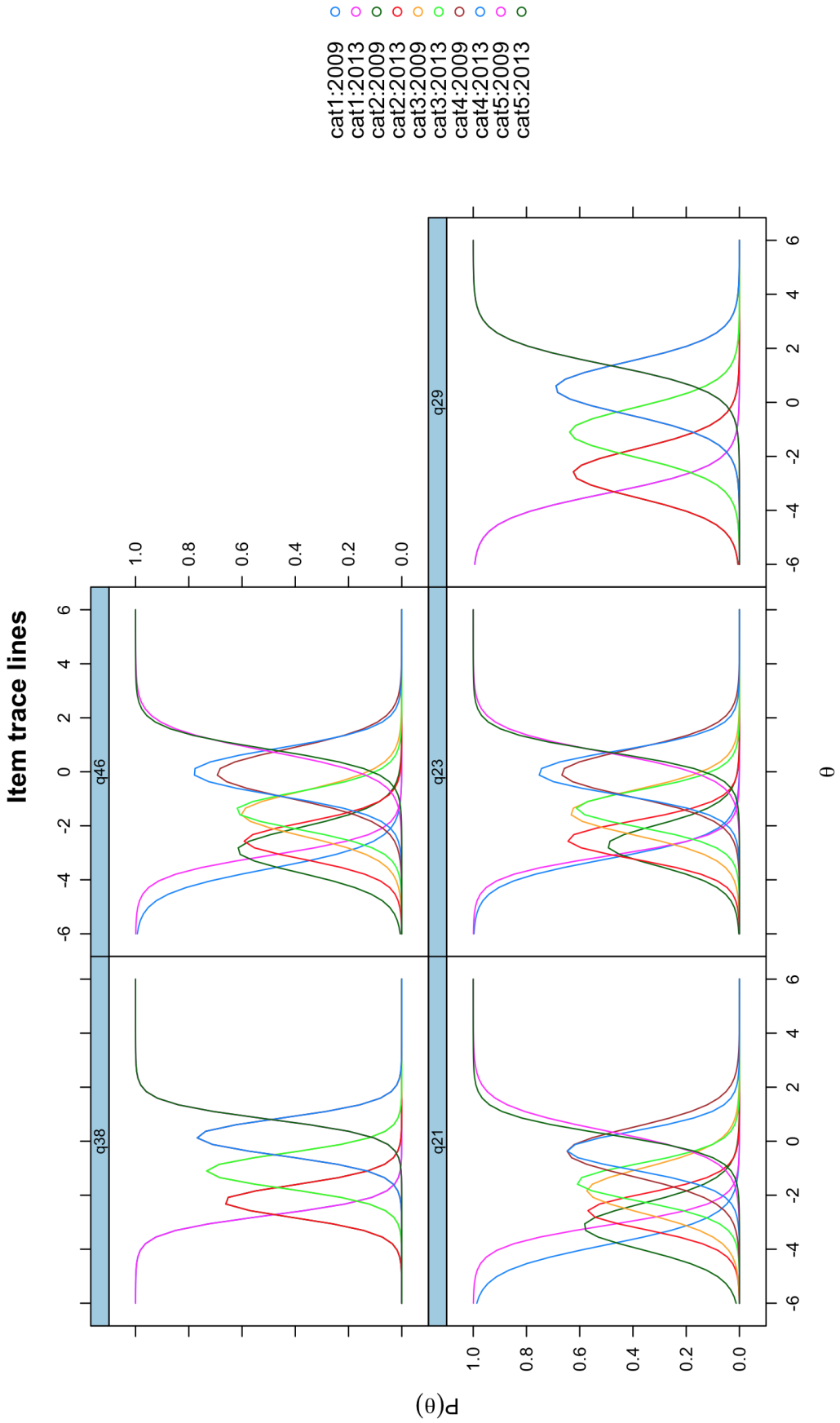
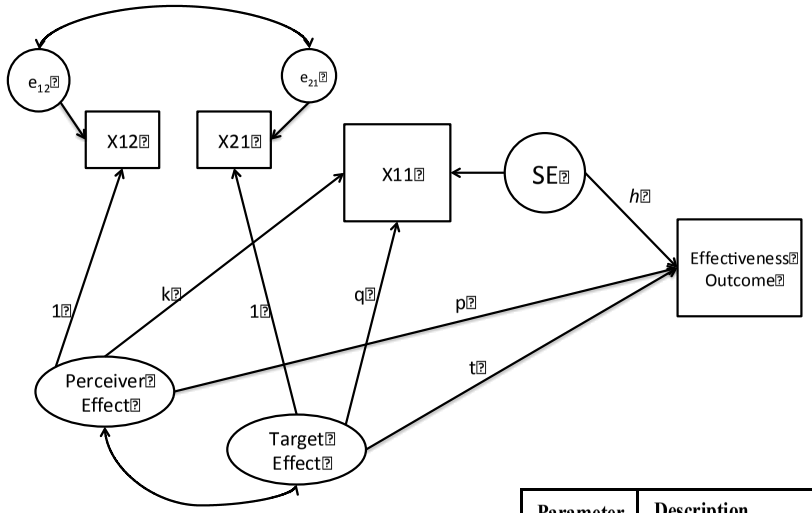


Figure 10

Socio-emotional DIF – Direct Report Ratings

Figure 11

Excerpt from Structural Equation Model Measuring Effect of Perceiver Effect, Target Effect, and Leader Self-Enhancement on Leader Effectiveness Outcomes



Parameter	Description
X12	Leader 1's rating of Leader 2
X21	Leader 2's rating of Leader 1
k	Effect of perceiver effect on leader self-rating; social comparison index
q	Effect of target effect on leader self-rating; self-insight index
p	Effect of perceiver effect on leader effectiveness outcome
t	Effect of target effect on leader effectiveness outcome
h	Hypothesized relationship between variance in leader self-ratings and leader effectiveness outcomes