PREDICTING WITHIN-SOURCE AGREEMENT IN MULTISOURCE FEEDBACK RATINGS: AN EXAMINATION OF CHARACTERISTICS OF THE RATER GROUP AND THE FOCAL MANAGER

by

CHRISTINE SCHRADER FERNANDEZ

A dissertation submitted to the Graduate Faculty in Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy,
The City University of New York

2008

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

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

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

This manuscript has been read and accepted for the Graduate Faculty in Industrial and Organizational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Dr. Karen Lyness

July 16, 2008
Date

Chair of the Examining Committee

Dr. Joseph Glick

July 16, 2008
Date

Executive Officer

Dr. Charles Scherbaum

Dr. Harold Goldstein

Dr. Walter Reichman

Dr. Allen Kraut

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

PREDICTING WITHIN-SOURCE AGREEMENT IN MULTISOURCE FEEDBACK RATINGS: AN EXAMINATION OF CHARACTERISTICS OF THE RATER GROUP AND THE FOCAL MANAGER

by

Christine Schrader Fernandez

Advisor: Professor Karen S. Lyness

Multisource feedback (MSF) involves gathering information about a manager's effectiveness from his or her boss, peers, and subordinates. Researchers typically average MSF ratings within rating sources (e.g., peers or subordinates), which assumes that agreement within rating sources is relatively high. However, there is little prior MSF research that has addressed the issue of within-source agreement, and the extant studies have often used inappropriate statistical techniques such as reliability indices. Moreover, this research often focuses on assessing the mean level of agreement or reliability within rating sources but has ignored the variability surrounding these indices. The purpose of the present study was to identify the predictors of agreement among peer and subordinate rater groups. Based on Kenny's (1991) weighted-average model of consensus, it was hypothesized that within-source agreement would be higher (1) for groups with higher levels of acquaintance with the focal manager, (2) for groups that were less diverse in terms of gender, race, age, and education, (3) for peers rather than subordinates, (4) for rating dimensions that raters have a high opportunity to observe rather than a low opportunity to observe, (5) for focal managers who are more extraverted, agreeable, and

conscientious, and (6) for focal managers who are more effective as rated by their supervisors. These hypotheses were tested with data from 33,696 focal managers who participated in the Benchmarks® multisource feedback program. The results indicated that peers had higher agreement than subordinates. Also, there were higher levels of agreement associated with more effective managers than less effective managers. Agreement was measured using a_{wg} and r_{wg} indices. These two indices were highly similar. However, a_{wg} was only calculated for about one-third of rater groups because many groups were too small or the group mean was outside of the interpretable range. The implications of eliminating groups are discussed. About three-quarters of peer groups and almost two-thirds of subordinates had high levels of agreement, however, an average of 5% of peers and 10% of subordinate groups failed to agree with one another at an acceptable level. The relevance of within-source agreement for MSF administration and feedback are discussed.

ACKNOWLEDGMENTS

My graduate education has been a long and arduous process at times. However, in looking back over the years, I feel incredibly fortunate to have had the opportunity to learn from some of the greatest minds in our field. My advisor, Karen Lyness has been incredibly supportive throughout the graduate school and the dissertation process. I am greatly appreciative of her insightful comments and suggestions that helped to shaped this dissertation. My dissertation committee members, Charles Scherbaum, Harold Goldstein, Allen Kraut, and Walter Reichman served as mentors throughout my graduate studies and I feel honored that they each took the time to provide thoughtful suggestions that helped refine this dissertation.

I would like to thank my family, my parents in particular, for encouraging me to face new challenges. I can't imagine having parents any more dedicated or proud.

Although graduate school helped me fulfill many of my scholastic goals, I think the biggest gift of graduate school was completely unexpected. I met my husband in the elevator of Baruch during my first week of classes. He first shared his teaching notes with me, and later acted as a sounding board as I honed my research questions, and provided much-needed pep talks when I was down. John, thank you for providing me with balance in my life. Your ability to calm me down at the end of the day brought fun, laughter and perspective into my life even when the big "D" was looming heavily over me.

I would like to thank the Center for Creative Leadership for providing the data used in this research.

Table of Contents

List of Tables

Chapter 1: Introduction of problem

Multisource feedback (MSF) or 360-degree feedback programs are a popular method of managerial evaluation and development in organizations (Chappelow, 2004; Fletcher & Baldry, 1999). Organizations use MSF systems to gather information from multiple perspectives regarding a manager's effectiveness on job-relevant skills. Specifically, superiors, peers, and subordinates individually rate a focal manager, and these ratings are compared with the manager's self-ratings. It is expected that this multisource feedback will provide a more comprehensive assessment of the manager's strengths and weaknesses than ratings from just one superior. This method of gathering multiple perspectives of feedback is especially valuable for managerial jobs which tend to be complex (Borman, 1974; Klimoski & London, 1974; Latham & Wexley, 1982; Tsui & Ohlott, 1988).

After ratings from multiple raters are gathered, the data are typically averaged within rating sources. For instance, after a group of five peers individually rate the focal manager, their ratings are combined to form an average peer rating score. This averaged score is often computed to protect the anonymity of individual raters and to simplify feedback (Fletcher & Baldry, 1999). In addition, researchers typically average scores within rating sources to evaluate the similarity of self-ratings with other-ratings, which is also known as self-other agreement (e.g., Fleenor, McCauley, & Brutus, 1996; London & Wohlers, 1991; Ostroff, Atwater, & Feinberg, 2004). Regardless of purpose, the result is an aggregated score that represents the mean rating within each rating source.

There are two underlying assumptions regarding MSF ratings when they are averaged within rating sources. The first assumption is that rating sources are likely to differ in their ratings (Borman, 1997; Bozeman, 1997; Klimoski & London, 1974). One possible explanation for differences among rater groups is that different rating sources have unique perspectives of the ratee (Borman, 1991; Landy & Farr, 1980; Murphy & Cleveland, 1995). In addition, it has been suggested that rater groups have different opportunities to observe the ratee (Lance & Teachout, 1992; Murphy & Cleveland, 1995), different definitions or implicit theories of job performance (Borman, 1974; Murphy & Cleveland, 1995; Tsui & Ohlott, 1988), and different rating motives (Conway, Lombardo, & Sanders, 2001; Murphy & Cleveland, 1995).

Regardless of the reason for the differences between rating sources, a number of research studies have investigated whether different rating sources provide unique ratings (e.g., Conway & Huffcutt, 1997; LeBreton, Burgess, Kaiser, Atchley, & James, 2003) and the findings from this body of research are mixed. For instance, a meta-analysis of the relationships between different groups' ratings found relatively low correlations among groups (e.g., $r = .34$ for supervisors and peers, $r = .22$ for supervisors and subordinates, and $r = .22$ for peers and subordinates; Conway & Huffcutt, 1997). Moreover, another meta-analysis found that both peer and subordinate ratings accounted for incrementally more variance in objective performance measures than supervisory ratings alone (Conway et al., 2001). In contrast, research that used generalizability theory and standards of agreement (Greguras & Robie, 1998; LeBreton et al., 2003; Scullen, Mount, & Goff, 2000) found fewer differences between rating sources and more variance associated with individual raters. Taken together, there is mixed support that each rater

group provides a unique perspective. Differences in sample characteristics are unlikely to be the reason for these different findings because most samples were large and spanned a number of organizations and industries. However, differing methodologies for calculating rating similarity may partially explain some of the mixed research findings, and such a possibility is examined by the present research.

The second assumption underlying MSF is that ratings within a given rating group are relatively similar (Borman, 1997; Bozeman, 1997; Van Velsor, Taylor, & Leslie, 1993). Many researchers suggest that MSF ratings from the same source should be similar because group members interact with the focal manager in equivalent settings and situations (Borman, 1997; Carless, Mann, & Wearing, 1998). For this reason, higher levels of convergence should occur within rating sources than between them. In general, there is some support for this assumption, particularly within the multitrait multirater literature (e.g., Borman, 1974; Kavanaugh, MacKinney, & Wolins, 1971). However, some recent research has failed to find high similarity in ratings within rating groups (e.g., Greguras & Robie, 1998; Mount, Judge, Scullen, Sytsma, & Hezlett, 1998; Scullen et al., 2000). For instance, Scullen at al. (2000) found that considerably more variance was attributable to the idiosyncrasies of individual raters than to rating sources. Again, it is possible that different research methods and analytical techniques may be the reason for differing results.

These mixed research findings are troubling as most researchers and practitioners would argue that having convergence within rating sources is important for a variety of reasons. First, the effectiveness of the feedback may decrease when raters do not agree in their ratings. If a manager receives feedback in the form of an averaged response, but

there was little underlying similarity among raters, the feedback will not effectively convey the separate opinions within the rater group that will ultimately guide the improvement of the manager's skills. In addition, London and Smither (1995) speculated that when managers receive information about the variability of their subordinates' ratings, they may be more apt to dismiss highly divergent ratings as being products of idiosyncratic ratings, than when there is a high level of convergence in ratings.

Agreement among raters is also important from a statistical standpoint in that it may not be statistically justifiable to collapse ratings within a rating source when raters make dissimilar ratings (Chan, 1998; James, Demaree, & Wolf, 1993; London & Smither, 1995; Mount et al., 1998). Composition models that combine individual-level data to express the perceptions of the group are only appropriate when ratings are similar to one another (Chan, 1998; Klein, Conn, Smith, & Sorra, 2001). For instance, if two peers rate a manager high on self-awareness, but two other peers rate the same manager much lower, an averaged feedback score will not accurately convey the groups' ratings, and therefore is not justified.

Finally, a high level of agreement is also desirable because agreement is often used as a proxy for the accuracy of the ratings (e.g., Borman, 1997). Researchers and practitioners associate lack of convergence with rating errors, suggesting that such ratings are less accurate than ones that are highly consistent among raters (e.g., Borman, 1997; Byrne, London, & Griffitt, 1968; LeBreton et al., 2003). Thus, reliability and agreement are often seen as necessary, but not sufficient conditions to have valid or accurate ratings (Cureton, 1951; Mount et al., 1998; Pedhazur & Schmelkin, 1991). From a classical test theory perspective, an individual's underlying true score cannot ever be assessed

(Crocker & Algina, 1986). Moreover, methods for estimating a ratee's true score in organizations are problematic because objective criteria are often inadequate and subjective ratings, even from trained experts, can be biased (Guion, 1998; Sulsky & Balzer, 1988). For these reasons, agreement is often used as a proxy measure for the ratee's true score on a particular trait (Wherry & Bartlett, 1982).

However, there are some situations where high agreement may actually indicate less accuracy (e.g., Schmitt, Noe, & Gottschalk, 1986). For example, rating response biases such as halo or leniency errors, can yield highly similar ratings among raters yet do not always accurately reflect an individual's actual behavior (e.g., Phillips & Lord, 1986). Stereotypes or implicit theories may also cause raters to agree with one another but fail to accurately capture the behavior of the focal manager (e.g., Lord, De Vader, & Alliger, 1986; Schmitt et al., 1986). In addition, it is important to note that there are situations where one may not expect raters to agree with one another, such as in the case of rater groups that have highly differentiated leader-member exchange relationships with the focal manager (London & Wohlers, 1991).

Thus, the level of convergence in rater groups may impact the effectiveness of feedback and has implications for the aggregation of individual-level data to group-level data. Convergence also has implications for assessing the quality of ratings. The present research takes the perspective that the relationship between within-source agreement and rating quality is complex. In some instances, high agreement may suggest accuracy and in others, bias. Of course, the problem of not being able to obtain a focal manager's underlying true score on a rating dimension only further compounds this problem. In an attempt to better understand what characteristics relate to within-source agreement, I

based my research hypotheses on Kenny's (1999) weighted-average model of consensus and accuracy. Kenny's mathematical model is based on prior work in the person perception literature (Anderson, 1981) and provides a theoretical rationale for how characteristics of the rater group and ratee relate to within-source agreement.

Summary of purpose

MSF programs are a popular way to gather feedback from a diverse set of constituencies. The underlying assumption is that multiple raters will provide a more comprehensive and more accurate assessment of a focal manager, which will subsequently help improve the individual's performance (London & Smither, 1995), and thus, the effectiveness of MSF programs rests on collecting accurate feedback. Because directly assessing the accuracy of individual ratings is impossible (Guion, 1998; Sulsky & Balzer, 1988), less direct methods such as examining the convergence of ratings are required. Such an investigation is complicated in that high agreement may indicate accuracy in some cases but not in others.

The present research examined the relationships between characteristics of the rater group and ratee to within-source agreement. This research was guided by Kenny's (1991) weighted-average model of consensus, which identified conditions when raters' assessments of individuals are more apt to be in consensus with one another. However, the present research extended this model of rating consensus to the MSF rating context. Moreover, the present research used a new measure for assessing within-source agreement (i.e., a_{wg}; Brown & Hauenstein, 2005), that had not been applied to MSF research. The present research was guided by the following questions: What types of rater groups tend to have a high level of consensus in rating a manager? Do characteristics of

the focal manager relate to increased within-source agreement? Are certain types of rating dimensions associated with higher levels of within-source agreement?

The following chapter provides a critical review of research that is relevant to examining the convergence of MSF ratings within rating sources and the possible predictors of this convergence. To start, I discuss the importance of accuracy in MSF ratings and the possible relationships between accuracy and convergence. Next, I review indices of rating similarity, including the differences between reliability and agreement and why the use of agreement is more appropriate for the purpose of examining MSF ratings. I also briefly review the findings from this literature, including studies that have used multitrait-multimethod studies or Pearson correlations (e.g., Kavanaugh et al., 1971; LeBreton et al., 2003; Mount et al., 1998), generalizability theory (Greguras & Robie, 1998; Webb, Shavelson, Kim, & Chen, 1989), and intraclass correlations (e.g., Atwater, Ostroff, Yammarino, & Fleenor, 1998; Fleenor, Fleenor, & Grossnickle, 1996; LeBreton et al., 2003; Ostroff et al., 2004). I conclude this section by reviewing the few studies that have specifically examined agreement within rating sources (Fleenor, et al., 1996; LeBreton et al., 2003); I also recommend the best index of convergence for use as the dependent variable in the present study. The literature review is generally confined to managers to reflect the target population of the present research; however, I include other populations when relevant.

Furthermore, I discuss the importance of studying the predictors of within-source agreement. This discussion includes why the adoption of a dispersion composition model (Chan, 1998) advances our understanding of MSF ratings. Then, I introduce a framework for identifying possible predictors of agreement within rating sources. This framework

integrates theory of performance rating (e.g., Landy & Farr, 1980; Murphy & Cleveland, 1995; Tsui & Barry, 1986) with Kenny's (1991) weighted-average model of consensus. Specifically, I examine how the characteristics of the rater group, the focal manager, and the dimension being rated may impact within-source agreement. Finally, I review the empirical support for each predictor of agreement to support my research hypotheses.

Chapter 2: Literature Review and Hypotheses

Accuracy and Convergence

Multisource feedback should provide focal managers with a candid snapshot of how they are viewed by their colleagues by uncovering their strengths as well as areas for development (Chappelow, 2004). Consequently, the success of MSF programs partially rests on the accuracy of the feedback given. Ratings are said to be accurate when they correspond to another set of measures, which are often referred to as true scores (e.g., Guion, 1998; Sulsky & Balzer, 1988). According to classical test theory, true scores are not attainable (Crocker & Algina, 1986). Thus, methods of approximating the underlying true score for job performance are necessary, and these often include using expert raters or averaging ratings (usually from trained experts; Sulsky & Balzar, 1988). For instance, novice raters' accuracy can be assessed by comparing their ratings to those of a trained expert. Because performance ratings are inherently subjective, however, even experts are susceptible to rating errors. For instance, Cronbach (1955) discussed how rater accuracy can diminish as a function of rating style (e.g., leniency) and cognitive processes such as stereotypes.

Because there are challenges in obtaining accurate ratings of job performance, convergence among raters is another frequently used proxy measure for accuracy (e.g., Borman, 1997; Van Velsor & Leslie, 2001). Convergence occurs when raters make comparable ratings for an individual, and is assessed using indices of reliability or agreement. Reliability provides information about the consistency of raters within a rating source, whereas methods of assessing agreement can additionally detect whether raters assigned the same rating value to the focal manager (Kozlowski & Hattrup, 1992).

The present research is specifically interested in agreement among raters; the reason for this decision and the specific computations will be discussed below. However, in this section, both reliability and agreement are discussed as they pertain to the relationship between accuracy and convergence.

When raters are in agreement in a MSF context, it suggests that the raters agree with one another about the strengths and weaknesses of a target. In addition, this convergence among raters also indicates that their ratings accurately assess the performance of the focal manager. For this reason, researchers stress the importance of obtaining relatively high within-source convergence as evidence that the ratings are accurate (e.g., Borman, 1997; Van Velsor & Leslie, 2001). However, accuracy and convergence may not always be positively related as is typically assumed (e.g., Borman, 1997). What if the group of raters all applied the same racial stereotypes when making their ratings? In this case, their ratings would be in agreement but would not be accurate. Thus, agreement and accuracy may not always be positively related. In this section, I will present two possible relationships between rater accuracy and convergence which differ from the classic assumption that convergence implies accuracy. Elaborating on these relationships is important for the present study because high levels of convergence among MSF raters may not always indicate accurate ratings.

Because similarity in ratings between raters is often used as a proxy for assessing rating accuracy, it is often deemed a desirable quality. However, a higher than expected level of similarity in some cases may indicate bias. Schmitt, Noe and Gottschalk (1986) used Brunswik's (1952) lens model to demonstrate this possibility. The key processes in the lens model are the way that raters synthesize their observations of a ratee to arrive at a

particular rating and the consistency with which they apply this synthesis in their overall ratings. The outcome of this model is the predicted level of reliability between two raters, which is estimated by combining the matching index and the consistency index. The matching index estimates the extent to which two raters combine rating dimensions similarly and is computed by correlating the two raters' predicted values of overall performance ratings with one another. The consistency index estimates the extent to which a rater consistently combines individual performance dimensions to arrive at an overall performance rating across ratees. This index is assessed by correlating the regression weights for overall performance ratings to determine the extent that each performance dimension consistently contributes to an overall performance rating. Thus, according to the lens model, if the predicted level of reliability is lower than the actual value, bias may be the reason. Bias or the use of information unrelated to performance, such as shared stereotypes, may actually inflate reliability or agreement because raters are applying similar information that is not the synthesis of performance dimensions. In this situation, high levels of interrater reliability could thus result in low validity, or accuracy in ratings.

Schmitt et al. (1986) used this model to examine the interrater reliability of job performance ratings for school administrators made by teachers and supervisors. After computing the estimated interrater reliability based on the calculated matching and consistency indices, the authors correlated demographic variables that were possible sources of rating contamination with the residuals of the ratings. If these demographic characteristics related to higher levels of interrater reliability one could infer that shared stereotypes, rather than the ratee's actual performance, influenced ratings. Although they

did not find indications of bias in their sample, Schmitt et al.'s use of the lens model (Brunswik, 1952) is valuable in that it illustrated how high levels of reliability can be undesirable when reliability relates to a characteristic that should not be related to job performance. Thus, the lens model provides one way to assess whether high levels of agreement or reliability may indicate bias rather than accuracy.

There are some limitations to the lens model, including the assumption that raters use the same rating dimensions to arrive at an overall job performance rating. Despite this major limitation, Schmitt et al.'s (1986) concepts are worth noting. Most important, this model raises the counterintuitive notion that high levels of interrater reliability or agreement may be a function of shared stereotypes rather than rating accuracy. Similarly, theory and research suggest that 'folk theories' (Borman, 1983; 1987) and implicit leadership theories (ILTs) (e.g., Foti & Lord, 1987; Lord, 1985; Lord, Foti, & de Vader, 1984; Phillips & Lord, 1986), may have a similar impact. Folk theories, Borman (1983; 1987) explains, are the idiosyncratic theories that individuals hold about what makes a person effective in a particular role. For instance, a rater may feel that impeccable presentation skills are necessary for a manager to be effective. Similarly, ILTs are individual's cognitive schemas and assumptions about the characteristics of the ideal leader (e.g., Foti & Lord, 1987; Lord et al., 1984). In both cases, raters use their idiosyncratic categories or schemas to guide their ratings. One implication is that these categories can lead to biased ratings if they overshadow the ratee's actual behavior or contain irrelevant characteristics (for review, see Phillips & Lord, 1986). Thus, raters using similar ILTs to make ratings are likely to be in agreement, but their ratings may not necessarily be accurate.

Both the lens model and ILTs imply that high agreement does not necessarily indicate accuracy, which has implications for the present research. For example, if raters within a group all hold the ILT that extraverted managers are more effective than introverted managers, they are likely to have high within-source agreement when rating a ratee. However, the high level of agreement may not necessarily indicate accurate ratings if the ILT does not accurately capture managerial performance.

The above scenario was one in which a high level of agreement was not necessarily desirable. There could also be situations in which a lack of agreement between raters may also be justifiable. Earlier I presented the idea that reliability is considered a necessary but not sufficient condition for accuracy (e.g., Cureton, 1951; Pedhazur & Schmelkin, 1991), suggesting that accurate ratings must also be reliable among raters. For this relationship between accuracy and agreement to be valid, raters must base their ratings on identical or fixed information and interpret this information in the same manner. However, some researchers have discussed the possibility, particularly in the person perception literature, that the relationship between the rater and the ratee is not fixed (Kenny & Albright, 1987; Swann, 1984). Specifically, the ratee is likely to display different behaviors based on the rater with whom he or she is interacting. One consequence is that individual raters may be justified in rating the ratee differently. For instance, managers may alter their leadership style to better match the needs of a subordinate. Also, focal managers may alter their behavior with different colleagues in an attempt to engage in impression management (London & Smither, 1995). The implication of these situations is that raters may differ in their rating of a focal manager, but that

these differences may accurately reflect each rater's unique relationship with the manager.

The unique relationship between a rater and ratee can also impact MSF ratings. For instance, London and Wohlers (1991) found a high level of variance in subordinates' ratings and speculated that the quality of leader-member exchange (LMX) relationships (e.g., Dansereau, Graen, & Hage, 1975; Graen & Uhl-Bien, 1995) may have been the cause. In this case, variability in ratings could accurately reflect each individual's appraisal of the leader's behavior. A member of the leader's in-group would probably rate the leader favorably on dimensions such as providing personal development. However, a member of the leader's out-group, who probably does not receive as much personal development from his or her supervisor, would probably rate the leader lower on this dimension. In this situation, the subordinates' ratings would not agree yet they would both reflect an accurate appraisal of the leader's behavior. Other factors may additionally impact the relationship between the rater and the ratee. For instance, research on relational demography suggests that individuals develop higher quality and more positive relationships with more similar individuals (e.g., Tsui & O'Reilly, 1989) and one consequence of this tendency is that groups that are heterogeneous are probably less likely to agree in their ratings of a ratee than a more homogenous group. This lack of agreement, however, may not reflect a lack of accuracy in the ratings; instead it may accurately reflect the unique impressions that group members hold toward the focal manager based on the quality of their relationship.

These two examples, one in which agreement indicates possible biases, and the other in which disagreements reflect valid differences, help to illustrate the complex

relationship between rater agreement and rating accuracy. Although it is not possible to determine with certainty whether agreement is a product of accuracy or bias in a field setting, such as the case of MSF ratings, the current study used a framework that provides theoretical explanations about how rater agreement relates to consensus and accuracy. However, before elaborating on the predictors of agreement, I will provide a discussion of the level of convergence typically found in MSF contexts. Specifically, I will present various methods for expressing rater convergence and discuss the extent of convergence typically found in MSF contexts.

Rating convergence within rating sources

One main assumption underlying multisource feedback, that is central to the present research, is that ratings within rating groups should be relatively similar (Borman, 1997). However, researchers have come to conflicting conclusions when examining the extent that raters within the same source provide similar ratings. Some researchers feel that adequate within-source similarity in ratings exists (e.g., LeBreton et al., 2003) and others do not (e.g., Greguras & Robie, 1998; Mount et al., 1998). Before examining the results of such studies, it is first important to examine the different ways in which convergence within rating groups has been examined.

Research on performance ratings has used two different standards to assess convergence within rating groups: reliability and agreement. Often these two standards are used interchangeably and it is possible that most of the disagreement regarding within-source similarity of MSF ratings is a by-product of using different statistical methods to measure this similarity. Most prior research used various methods to assess the reliability of ratings within rating sources (e.g., Greguras & Robie, 1998; Mount,

1984; Mount et al., 1998). Methods that measure reliability provide information about the consistency of raters within a rating source, but do not assess whether raters are in agreement, or assign the same numerical ratings to the focal manager. Less frequently, research has specifically assessed the agreement, or whether raters provide the same ratings within rating sources (e.g., Fleenor, Fleenor et al., 1996; LeBreton et al., 2003). Because understanding the differences between reliability and agreement is essential to understanding the current research, I first discuss the differences between reliability and agreement and then review the research that has examined the similarity of MSF ratings using these different types of standards.

Reliability and agreement. The terms "interrater reliability" and "interrater agreement" are often used interchangeably when discussing MSF ratings, yet these two types of indices provide different information about the extent that two raters' ratings are similar. This divergence is because reliability indices assess the consistency of ratings across raters, whereas agreement indices can additionally detect whether different raters assign different values to their ratings (Kozlowski & Hattrup, 1992). One implication is that raters that are consistent need only agree in their rank ordering, whereas raters need to give the same rating values to be in agreement.

It is important to note that reliability and agreement are not necessarily related, but rather, can be quite independent of one another. One illustration of this notion is when raters consistently rate an individual yet fail to assign similar values to the ratee. Take for example two raters rating a focal manager. If the first rater rates the manager a 3, 4, and 5 on dimensions 1, 2, and 3, and a second rater gives ratings of 1, 2, and 3 for the same dimensions, we would find that the raters have high interrater reliability. In

contrast, we would not find high agreement because they did not give the same rating value for any of the dimensions.

The opposite scenario, having high agreement but low reliability, is also possible when there is range restriction among raters across ratees. For instance, if two peers rate eight focal managers on a 5-point scale (Peer 1: 2, 2, 1, 1, 5, 5, 4, 4; Peer 2: 1, 2, 1, 1, 4, 5, 4, 4), one could calculate both the reliability and the agreement of their ratings (LeBreton & Senter, in press). In doing so, there would be high interrater reliability (r = .96) *and* high interrater agreement (r_{wg} = .97; the r_{wg} agreement index will be discussed in greater detail later in this paper). However, the values of these two indices are less related in situations of range restriction. To demonstrate the impact of range restriction, reliability and agreement could be calculated for the first four and last four ratings separately. In this example the reliability would be lower (r = .58) whereas agreement would remain high (r_{wg} = .97). Kozlowski and Hattrup (1992) explained that having restriction in range decreases power in indices that assess response consistency because the reliability indices are essentially calculating consistency with a compressed scale rather than for all five values of the response scale. Therefore, it is possible to have high agreement but low reliability in situations with range restriction.

The issue of agreement versus consistency is also important when considering the hypotheses that one wishes to make. In most prior research, reliability has been used to assess similarity within rating sources. Although these studies do provide some answers to questions of rating similarity, Kozlowski and Hattrup (1992) recommend using agreement standards, rather than reliability standards, when one is aggregating responses or using composition models. Composition models combine data at one level of analysis

to represent a higher level of analysis (Chan, 1998; Klein, Dansereau, & Hall, 1994). To illustrate, individual MSF ratings represent the individual-level of analysis because each rating represents one individual's perspective. Yet, if we wanted to capture how the group rated a particular ratee, we could average individuals' ratings within the group. In this case, the averaged score would represent the group-level of analysis. However, an averaged score will only adequately reflect the perceptions of the group when there is high agreement within the group. Thus, the level of rater agreement is important when applying a composition model or aggregating data.

Aside from the issue of aggregation, agreement indices are more appropriate than reliability indices when the actual rating value is important (Guion, 1998; Kozlowski & Hattrup, 1992). For instance, the MSF score that a rater assigns to the ratee conveys the quality of the focal manager, such that the difference between receiving a rating of 4 and a 5 is likely to be meaningful, assuming that ratings are made on an interval scale. Because agreement is able to reflect the similarity of actual values given to a ratee by multiple raters, it is agreement, and not reliability, that should be the standard when examining within-source convergence.

Although agreement does seem to be the most appropriate type of index to use when studying the similarity of ratings within rating sources, the bulk of prior research in this area has employed standards of reliability, through the use of Pearson correlations (e.g., Mount, 1984), confirmatory factor analysis (e.g., Mount et al., 1998), and generalizability theory analyses (e.g., Greguras & Robie, 1998). By reviewing these types of analyses, I will briefly summarize the main research findings and also discuss the weaknesses that are associated with each type of reliability index. In addition, ICCs

(intraclass correlations) have been used to provide information about within-source similarity (e.g., Fleenor et al., 1996). Often there is confusion concerning ICCs because they can be used to assess reliability and agreement depending on the equation used. I will explain the different forms of ICC that have been used to assess within-source similarity of MSF ratings and review studies that have used ICCs to assess within-source similarity. Finally, I will review and critique different measures that solely assess agreement among raters, including the standard deviation (SD_x), the T index, the Finn Index, r_{wg}, the average deviation (AD), and a_{wg}. I will also review in depth the few studies that have used standards of agreement to examine MSF ratings.

Multitrait-multimethod studies. One of the most frequent methods of assessing the reliability of raters within a rating group is the multitrait-multimethod (MTMM) approach (e.g., Kavanaugh et al., 1971; Mount et al., 1998; Scullen et al., 2000). Generally, this approach examines the variance in performance ratings that is associated with the dimension being rated and the method of rating, or the rating source (Campbell & Fiske, 1959). For instance, if relationships are high across sources, it would suggest a high degree of convergence among rating sources. In addition, if relationships are high among rating dimensions, it is typically thought to indicate a halo effect, or that raters are not making distinctions among rating dimensions. Moreover, relationships among rating dimensions can also be assessed in these types of studies to assess the factor structure of rating dimensions. There are a variety of statistical methods used in MTMM studies, including confirmatory factor analysis, correlated uniqueness, and the direct product model. Although an in-depth review of these methods is beyond the scope of this paper, Becker and Cote (1994) and Conway (1996) both provide useful discussions of these

issues. For the present paper, however, I will specifically focus on the reliability within different rating sources.

One approach to conveying the findings of MTMM studies is to report the Pearson correlation coefficients that express the strength of relationships between two variables. The average of the cross-product correlations can be used to estimate the extent that raters within the same source provide similar ratings (e.g., Kavanaugh et al., 1971; Mount et al., 1998). The average correlations usually indicate low levels of reliability within rater groups. For instance, Mount (1984) found that the average correlations among subordinates' ratings on various managerial behaviors ranged between .18 and .28. In addition, Mount et al. (1998) also found that the correlations within rater groups on the same traits on a developmental MSF instrument were quite low, ranging between .26 and .31 for peers and between .31 and .34 for subordinates. More recently, LeBreton et al. (2003) compared Pearson correlations within rating sources and found an average correlation of .30 for both peers and subordinates.

Together, the findings suggest that reliability is not particularly high within rating sources. However, it should be noted that these findings may partially result from the statistical artifact of range restriction (LeBreton et al., 2003). To demonstrate, the formula for Pearson correlations is:

$$r = \frac{s_{XY}}{s_X s_Y} \qquad (1)$$

where s_X and s_Y are the standard deviations of variables X and Y, respectively, and s_{XY} is the covariance of X and Y (Glass & Hopkins, 1996). Thus, when the values of a variable are relatively homogenous, the standard deviations of the variables will be low and the resulting correlation coefficient will be smaller than in situations with greater variability

(Alliger & Williams, 1989; Cohen, Cohen, West, & Aiken, 2003). LeBreton at al. argued that this type of restriction is likely to occur within MSF ratings because of a number of organizational processes such as recruitment, selection, and training help create high-performing employees, thus reducing the variability among ratees. Also, subordinates may be motivated to inflate ratings as a way to maintain favorable relationships with the focal manager (Murphy & Cleveland, 1995). Prior research has also found evidence that MSF ratings are restricted in range (e.g., Mount, 1984; Walker & Smither, 1999), particularly when the instrument is used to give administrative rewards (Greguras, Robie, Schleicher, & Goff, 2003). As a result, Pearson correlations may provide attenuated estimates of within-source convergence.

Attenuated estimates of reliability are problematic when assessing convergence among raters. Although methods for correcting range restriction are often used in performing meta-analyses and building personnel selection tools (e.g., Hoffman, 1995; Hunter & Schmidt, 1990; Sackett, 2000; Sackett & Ostgaard, 1994), doing so when examining the convergence of MSF ratings would be inappropriate. When assessing rater convergence, researchers are interested in understanding the extent to which raters agree with one another about a ratee, and for this reason, some degree of range restriction is expected and even desirable. Although range restriction may be unavoidable in such situations, its impact on correlation coefficients is important to understand.

Other MTMM methods using confirmatory factor analysis and correlated uniqueness have also found low levels of convergence within rating sources. For instance, Mount et al. (1998) performed confirmatory factor analysis (CFA) to examine the factor structure of MSF ratings to determine whether ratings were unique by rating

source. They determined the fit of a variety models using MSF ratings from two supervisors, two peers, two subordinates, and one focal manager. The seven-factor model that included a separate factor for each rater source (i.e., self, peer, subordinate, and supervisor) and each of the three rating dimensions (i.e., human relations, technical, and administrative skills) had a poor fit (RMR = .14; GFI = .67). However, the nine and ten-factor models had better fits. The nine-factor model, which was comprised of six rater factors (i.e., two peers, two subordinates, one self, and supervisors combined) and the three rating dimensions (i.e., human relations, technical, and administrative skills), had relatively good fit (*RMR* = .13; *GFI* = .96). The ten-factor model had even better fit (*RMR* = .02; *GFI* = .99) and was similar to the nine-factor model except the two supervisor ratings were treated as separate sources. These results suggest that at best, only supervisor ratings have enough similarity to comprise a rating source. Put another way, ratings within peer and subordinate groups were not similar enough to form a unique rating source.

Scullen, Mount, and Goff (2000) investigated the latent relationships in MSF ratings among supervisors, peers, subordinates, and self ratings using correlated uniqueness and CFA methods. Their research specifically examined how a ratee's general job performance, a ratee's job facet performance (i.e. human, administrative, and technical skills), idiosyncratic rating tendencies, rating source, and measurement error impacted MSF ratings. When they partitioned variance according to these five factors, they found that the ratee's general and facet performance only accounted for a combined average of 21 to 25% of the variance in performance ratings. In addition, they found that a majority of the variance could be attributed to idiosyncratic rating tendencies. The

researchers suggested that there are two main types of idiosyncratic errors. First, halo errors (Balzer & Sulsky, 1992; Thorndike, 1920) occur when ratings are influenced by a general assessment of a ratee rather than the rater's behavior on specific dimension that is being rated. Second, leniency errors (Guilford, 1954) are the tendency for raters to give systematically favorable (or unfavorable) ratings. However, Scullen et al. were not able to partition the variance attributable to each type of error.

Together the two studies that used CFA techniques found little indication of similarity within rating sources. In particular, Scullen et al. (2000) found an indication that individuals' ratings were fraught with idiosyncratic rating errors or biases. Because of this finding, they recommended that future research should investigate possible causes of these large idiosyncratic rater effects. However, as with other studies based on reliability, it is possible that range restriction may have impacted these findings of the above MTMM studies (LeBreton et al., 2003). In addition, it is important to note that halo and leniency errors that comprise idiosyncratic ratings can be confounded because rating magnitudes have been shown to relate to some indices of rating variability (Alliger & Williams, 1989). Therefore, the present research seeks to examine the predictors of rater group agreement by minimizing this confound. Specifically, I examine agreement in a manner that is less impacted by rating magnitude than prior research studies.

Generalizability theory. Generalizability theory, or G-theory, is another method for determining the reliability of MSF ratings (Cronbach, Gleser, Nanada, & Rajaratnam, 1972; DeShon, 2002; Murphy & DeShon, 2000; Shavelson & Webb, 1991). G-theory diverges from the classical test theory assumption that ratings are composed only of true scores and random error. Instead, G-theory is based on the idea that the error terms can be

further deconstructed into systematic and random sources of error using a single research study (e.g., error attributable to the rater, ratee, and dimension). One can examine the role that each source of error plays in explaining variance in scores, as well as the interaction between these sources. Once these variance components are estimated, researchers can additionally determine the number of raters needed to attain particular levels of reliability. G-studies are beneficial for these reasons, but they have limited utility with research designs that are not fully crossed (DeShon, 2002), as is the case with MSF ratings. Moreover, the present research requires a dependent variable that can express rater group variability relatively independent of agreement (Brown & Hauenstein, 2006). For these reasons, I will not be able to use a G-study for the present research. Thus, the specific steps for computing the variance components of a G-study are beyond the scope of the current paper (for an in-depth review, see DeShon, 2002; Shavelson & Webb, 1991); however, I will discuss some of the conclusions of within-source reliability that have been based on G-studies.

Three MSF studies used generalizability theory to examine the reliability of performance ratings. Kraiger and Teachout (1990) examined self, peer, and supervisor ratings of Air Force mechanics. They found each rater source to be reliable; however, only one rater was included in each source. Webb, Shavelson, Kim, and Chen (1989) similarly examined self, peer, and supervisor ratings of performance for Navy machinists but included ratings from two peers. The groups of two peers, they found, had relatively low reliability in the ranking of machinists. They estimated that five peers would be necessary to reach a generalizability coefficient of .80. The findings of Webb et al. suggest that ratings within rater groups may not be reliable. Aside from the conflicting

findings of these two studies, it is important to note that both used military populations in trade jobs and their findings may not apply to a managerial population.

A study by Greguras and Robie (1998), however, did use managers. Also, unlike prior generalizability studies, they gathered multiple ratings from supervisors, peers, and subordinates. They found that the Item and the Ratee x Item effects accounted for little variance in the ratings across all three groups; this finding indicates that the item variance does not have a large impact on the reliability of ratings. In addition, the interaction between specific items and ratees did not account for much variance. However, they did find a relatively large rater main effect combined with a Rater x Ratee effect for all groups, with the largest in subordinate groups. This finding suggests that there was little within-group reliability for all rating groups and that individual raters appeared to make idiosyncratic ratings. However, because the design was such that raters were nested within ratees, it is impossible to disentangle the variance attributable to the rater versus the interaction between the rater and ratee. Despite this limitation, Greguras and Robie were able to estimate the number of raters needed to attain a generalizability coefficient that exceeded .70. For a five-item measure nine subordinates, eight peers, and four supervisors are needed; eight subordinates, seven peers, and four supervisors are needed for a ten-item measure; seven subordinates, six peers, and four supervisors are needed for a twenty-item measure. Therefore, even with a twenty-item measure, the desirable number of raters per rating source exceeds the numbers that are usually used in most MSF programs (e.g., Chappelow, 2004).

These findings do suggest a lack of within-source consistency. However, Greguras and Robie found an even larger proportion of variance could be attributed to a

Rater x Item, Ratee x Rater x Item plus residual error that could not be further disentangled. To further understand that large Rater x Ratee effect found, they suggested that future research should examine other contextual factors, such as the rater's opportunity to observe the ratee, which may additionally explain variance in MSF ratings. In addition, they concluded that future research should examine whether characteristics (e.g., demographic, attitudinal, or rater-ratee relationship) of the raters and ratees help to explain the lack of convergence within rating sources. Their suggestions raise valid points and reflect the aims of the present study. By examining the extent to which variability in ratings is associated with characteristics of the rater and ratees, we can gain a better understanding of MSF ratings.

The study by Greguras and Robie provides useful information about the convergence of ratings within rating sources and the possible sources of this variance. However, in general, G-studies are limited in their utility because of the nested design (i.e., raters nested in ratees) that is typically used in MSF ratings. This design limits the extent that variance components can be partitioned (Brown & Hauenstein, 2005; Greguras & Robie, 1998). Also G-studies cannot provide individual estimates of agreement for each rating group which is required for the present research.

Intraclass correlations. Unlike interclass correlations (i.e., Pearson's *r*) that compare measures of different classes or metrics, intraclass correlation (ICCs) assess the relationships between variables that share common metrics and variance (McGraw & Wong, 1996; Shout & Fleiss, 1979). Specifically, in the case of MSF ratings, ICCs express the ratio of systematic variance which is attributable to the differences between ratees to the total variance within MSF ratings. Depending on the form used, ICCs can

assess both the consistency and agreement in situations with multiple raters (e.g., McGraw & Wong, 1996). ICCs that measure agreement (i.e., ICC(A)) differ from consistency measures (i.e., ICC(C)) in that they include the variance of the column terms in the denominator that represents the values that raters assign to each item (for a complete review of all ICC formulas, see McGraw & Wong, 1996). In addition, forms of ICCs can also differ based on whether the study is a one-way or two-way design. I will omit further discussion of the various forms of two-way designs, because, for the case of MSF ratings, a one-way random effects model is appropriate. This design is appropriate because each ratee is rated by a different set of raters and raters are nested within ratees (Fleenor et al., 1996; LeBreton & Senter, in press; McGraw & Wong, 1996; Shout & Fleiss, 1979). In addition, the ratee is typically treated as a random effect.

Although ICCs can measure both agreement and consistency, a one-way random effects model only assesses the agreement among raters. McGraw and Wong (1996) explain, "for one-way models there are no C-type coefficients because only absolute agreement is measurable in this context" (p. 34). Therefore, ICCs used to analyze MSF ratings only convey information about agreement, but not reliability. The notation for the one-way random effect ICC differs; Shout and Fleiss (1979) call this form ICC(1, 1), whereas McGraw and Wong (1996) use the term ICC(1). Using McGraw and Wong's notation, ICC(1) is calculated as:

$$\text{ICC}(1) = \frac{MS_R - MS_W}{MS_R + (k-1)MS_W} \qquad (2)$$

where, k is the number of raters, MS_R is the mean square for the row (the ratee), MS_W is the mean square for residual sources of variance. The resulting values of ICC(1) measure the agreement on a single measurement and can be interpreted as the anticipated level of

agreement of a single rater across ratees (LeBreton & Senter, in press). However, in instances where information about the average level of agreement of a group of raters is more important, the form ICC(k) is appropriate. This metric is also a one-way random effects model. In the case of the MSF ratings, ICC(k), which is also called ICC(1, k) according to Shout and Fleiss (1979), can assess the anticipated agreement and reliability of a group of k raters and is calculated as:

$$\text{ICC(k)} = \frac{MS_R - MS_W}{MS_R} \tag{3}$$

where MS_R is the mean square for the row (the ratee) and MS_W is the mean square for residual sources of variance. Researchers who calculated ICCs for MSF rating groups tended to find inconsistent results. For comparison, I will review research using ICC(1) and ICC(k) separately and refer to analyses using McGraw and Wong's (1996) notations. In addition, I will note if the original source used a different notation.

A few studies on the agreement of individual raters used the ICC(1) metric. For example, LeBreton et al. (2003), who used a large sample spanning multiple organizations and industries, found that the mean ICC(1), which they reference as ICC(1, 1), was .30 for subordinates and .31 for peers. In addition, Fleenor et al. (1996) found within a single health care organization that ICC(1), which they refer to as ICC(1, 1), for subordinate raters ranged between .04 and .34, with a median of .20. The reported values of ICC(1) provide information about the level of agreement found for an individual's ratings, and provide some information about the typical stability of individual raters.

In addition, ICC(k), which measures the agreement for k raters, has also been calculated for MSF rating sources to estimate the stability of ratings for a group of raters (McGraw & Wong, 1996). This form of ICC is particularly relevant for assessing the

agreement within rating groups. Typically, researchers express this form of ICC according to group size, such that ICC(1, 3) estimates the agreement of three raters. LeBreton et al. (2003) estimated the agreement of five raters from a large and diverse managerial sample. They found that the average ICC(1, 5) was .68 for peers and .67 for subordinates. Also using similar managerial samples, two additional studies estimated ICC (1, 3). One study reported that the agreement among subordinates ranged between .47 and .70 (M = .59) (Fleenor, McCauley, & Brutus, 1996); another study reported that ICC (1, 3) ranged between .43 and .69 for peer raters (Atwater et al., 1998). Note that Atwater et al. did not specifically reference the form of ICC used, but they attributed their method to Fleenor et al. (1996). Ostroff et al. (2004) calculated ICC(1, 3), which they called ICC(3), for a large sample of managers from a diverse group of organizations; the average agreement was .64 for subordinates and .61 for peers. In addition, Fleenor et al. (1996) in a study of a single organization, calculated ICC(1, 3) for subordinate raters; the values ranged between .12 and .61 with a median of .43.

The ICC values found in these various settings indicate that agreement is variable within rating groups, suggesting that it is important to further investigate these differences in levels of agreement. However, conclusions based on the ICC metric are clouded by a few issues. First, although McGraw and Wong (1996) attempted to standardize the naming conventions of ICC formulas, which differed from Shout and Fleiss's (1979) conventions, there still appears to be some ambiguity in how different forms of ICCs are reported. In addition, both ICC(1) and ICC(k), are measures of agreement (McGraw & Wong, 1996), but are sometimes discussed as a measure of reliability (cf. Fleenor et al., 1996). Also, some of the variability in the ICC coefficients

reported by researchers may be due to the sensitivity of ICCs to the variability of the sample (e.g., LeBreton et al., 2003)**,** or the extent that there is variance among ratees. Moreover, because values of ICCs are attenuated in samples with range restriction, or when ratings or levels of performance do not vary substantially across ratees, it is inadvisable to apply specific ranges of what constitutes a 'high' level of agreement.

Agreement

With the exception of ICCs, which can assess both reliability and agreement, the above methods use standards of reliability to assess the convergence of ratings from the same source. In contrast, there are a number of methods that are specifically designed to only assess agreement, or measure the extent that raters assign the same values to a ratee. These indices range from basic measures such as the standard deviation index (Schmidt & Hunter, 1989), to more complex measures such as the *T* Index (Tinsley & Weiss, 1975), the Finn index (1970), r_{wg} (James, Demaree, & Wolf, 1984; James et al., 1993; Lindell, Brandt, & Whitney, 1999), the average deviation index (AD; Burke, Finkelstein, & Dusig, 1999; Dunlap, Burke, & Smith-Crowe, 2003), and most recently, a_{wg} (Brown & Hauenstein, 2005). The following section provides a brief discussion of how indices of agreement have evolved (for a complete discussion of this evolution, see Brown & Hauenstein, 2005; Kozlowski & Hattrup, 1992). I confined my review of agreement indices to those that are appropriate for use with continuous variables in situations where a single stimulus is rated (Brown & Hauenstein, 2005). In addition, I discuss the calculations of the most widely accepted measures, and provide a detailed analysis of the strengths and weaknesses of these measures. Finally, I review the research that has assessed levels of within-source agreement using various measures of agreement.

One of the most basic measures of agreement is SD_X or, the standard deviation of the ratings of a ratee across raters (Schmidt & Hunter, 1989), which is calculated as:

$$SD_X = \frac{\left(X_k - \overline{X}\right)^2}{k-1} \quad (4)$$

where k equals the number of raters, X_k is the rater k's rating on X, and $\overline{X}$ is the mean rating across the k raters. Schmidt and Hunter additionally recommended calculating the standard error, SE_M, to construct 95% confidence intervals to estimate the error in the ratings,

$$SE_M = \frac{SD_X}{\sqrt{k}} \quad (5)$$

where SD_X is divided by the square root of k raters. It is important to note that this metric actually measures dispersion, or the extent to which there is variance among raters (Chan, 1998; Feinberg, Ostroff, & Burke, 2005; Klein et al., 2001), rather than agreement. As such, values increase with disagreement rather than agreement. A score of zero represents perfect agreement and lower levels of agreement are represented by larger values. For this reason, LeBreton and Senter (in press) suggested that this metric may be more appropriate for dispersion composition models. For example, SD_X has been used as a dependent variable in organizational climate research (Klein et al., 2001). However, this statistic has some shortcomings. Kozlowski and Hattrup (1992) argued that this metric fails to consider that some raters may agree based on chance rather than actual agreement, thus inflating agreement under conditions of response bias. In addition, the values of SD_X vary as a function of the number of raters (fewer raters relate to smaller values of SD_X), and also vary according to the number of response options. Also, unlike

other indices of agreement that range between zero and one, SD_X does not have a fixed interval, which further complicates interpretation.

Other measures of agreement evolved from indices that were aimed at assessing agreement on nominal variables and used proportion of agreement as the basis for computing agreement (e.g., Cohen's (1960) kappa). These indices were problematic for assessing agreement for continuous variables, such as performance ratings, because they express agreement in absolute terms (i.e., yes or no) and could not accommodate varying degrees of agreement (Kozlowski & Hattrup, 1992). Put another way, these early indices of agreement could capture the level of agreement for nominal categories, as in the case of diagnosing a patient's illness; however, they were problematic when raters have a greater number of response choices, as is the case with MSF ratings. Kozlowski and Hattrup (1992) noted that the second generation of agreement indices provided some improvements on earlier methods of agreement. These indices include those advanced by Lu (1972), Lawlis and Lu (1972), Tinsley and Weiss (1975), and Finn (1970) and were able to distinguish varying levels of agreement among raters.

Tinsley and Weiss (1975) proposed the *T* Index, which can account for agreement based on chance for a group of raters making ratings on an interval scale. The *T* Index is computed as:

$$T = \frac{N_1 - NP}{N - NP} \tag{6}$$

where N_1 is the number of agreements among raters, N is the total number of items rated, and P is the probability of chance agreement on an item. The probability of chance agreement is determined through the following formula (Lawlis & Lu, 1972):

$$P = \frac{(n-1)\sum_{i=1}^{k-1} 2^{k-1} + n}{n^k} \quad (7)$$

where k is the number of raters in the rating group and n is the number of points on the scale. Larger values of the T index represent a greater level of agreement. The range in values depends on the number of response choices, number of raters, and probability of chance agreement (Lindell et al., 1999). One of the advancements of the T Index is that researchers can specify the margin by which raters could differ and still be considered in agreement. An example is allowing raters to differ in their ratings by one point but still consider them to be in agreement. However, this method still distills agreement into a dichotomous process (Kozlowski & Hattrup, 1992).

Another method of assessing agreement is the Finn (1970) index, which assesses agreement for multiple raters judging a single ratee on an interval scale of measurement. This statistic is calculated by dividing the observed variance by the expected variance. The formula for the Finn Index is:

$$\text{Finn Index} = 1 - \frac{S_X^2}{\sigma_E^2} \quad (8)$$

where S_X^2 is the observed variance of the ratings for a ratee on item X across a group of raters, and σ_E^2 is the expected variance. The term σ^2_E is also referred to as σ^2_{EU} (James et al., 1984) to represent the expected variance of a uniform distribution. The expected variance of a uniform distribution, which is also called a rectangular distribution, assumes that raters are equally likely to select any of the possible response choices. The expected variance is calculated as follows (Mood, Graybill, & Boes, 1974):

$$\sigma_E^2 = (A^2 - 1)/12 \quad (9)$$

where A is the number of response options that raters can choose from. Thus, with five possible response options, the expected variance would be 2.0 ($\sigma_E^2 = [5^2 - 1]/12$). Higher values of the Finn Index correspond to higher levels of agreement, with an upper limit of 1.0 representing perfect agreement.

The Finn Index is thought to represent "the proportion of non-error variance in the ratings" (1970, p.72) such that agreement among raters occurs when the variance of their ratings is less that what would be expected by chance. One problem with this index is that the assumption of a rectangular distribution is unlikely to be valid because performance ratings tend to be negatively skewed (LeBreton et al., 2003; Mount, 1984; Murphy & Cleveland, 1995; Walker & Smither, 1999). The implication is that skewed ratings are apt to inflate estimates of agreement with the use of a rectangular distribution.

Because ratings are unlikely to be uniformly distributed in most rating conditions, James, Demaree, and Wolf (1984) adapted the Finn (1970) index to form r_{wg}. By varying the expected variance term of this index, researchers are able to account for the possibility that raters agreed due to chance or because of response bias. To express the agreement of raters on a single item X, the general formula for r_{wg} is:

$$r_{\text{wg}} = 1 - \frac{S_X^2}{\sigma_E^2} \qquad (10)$$

where S_X^2 is the observed variance for a ratee for item X across a group of raters, and σ_E^2 is the expected variance of ratings for item X. Similar to the Finn Index, r_{wg} is a ratio of the observed and expected variance and represents the proportional reduction in error variance. This measure is able to assess agreement for one-item measures. Values of r_{wg}

increase with levels of agreement. Values of r_{wg} range between 0 and 1.0 for dichotomous variables. With a greater number of response options, negative values can be obtained when the observed variance is greater than the expected variance. However, 1.0 remains the upper limit of agreement (Lindell et al., 1999).

One of the main features of the r_{wg} index is the ability for researchers to specify what the expected variance among raters should be. When there are theoretical reasons why raters may be influenced by social desirability or response biases, James et al. (1984) recommended using a null distribution to appropriately reflect the rating scenario. Specifically, researchers should specify the smallest and largest expected variances based on theory and prior research, and then calculate r_{wg} for those null distributions, and in essence, create a range of agreement values. To provide some guidance in selecting the correct null distribution, James et al. provided specific calculations for triangular (central tendency) and negatively skewed null distributions. In addition, LeBreton and Senter (in press) provided the expected variance terms for six different null distributions.

James et al. (1984) also developed $r_{wg(J)}$ to calculate agreement for scales with J essentially parallel items by substituting the average item variance for the observed variance and applying the Spearman-Brown prophecy. The Spearman-Brown formula illustrates that adding parallel items to a scale will increase reliability and is expressed as (Crocker & Algina, 1986):

$$\rho_{xx'} = \frac{k\rho_{jj'}}{1 + (\mathrm{k} - 1)\rho_{jj'}} \tag{11}$$

where $\rho_{jj'}$ is the reliability of a single test, and k is the number of items in the scale and $\rho_{xx'}$ is the predicted reliability of the scale with k items. The resulting $r_{wg(J)}$ statistic is

calculated as:

$$r_{\mathrm{wg(J)}} = \frac{J\left(1 - \frac{\overline{S}^2_{X_j}}{\sigma^2_{\mathrm{E}}}\right)}{J\left(1 - \frac{\overline{S}^2_{X_j}}{\sigma^2_{\mathrm{E}}}\right) + \left(\frac{\overline{S}^2_{X_j}}{\sigma^2_{\mathrm{E}}}\right)} \tag{12}$$

where, $\overline{S}^2_{X_j}$ is the mean of the observed variances for a particular ratee across raters computed for J essentially parallel items, and σ^2_E is the expected variance of some specified null distribution. One implication of applying the Spearman-Brown prophecy to calculations of $r_{wg(J)}$ is that the resulting agreement coefficients for a scale are higher than the agreement for the individual items (LeBreton & Senter, in press).

For both the single-item and scale forms of the r_{wg} statistic, the main advancement over the Finn (1970) index is that researchers are able to specify null distributions other than the uniform distribution, depending on the rating scenario. This feature is particularly pertinent when raters are not expected to endorse each response choice equally. LeBreton and Senter (in press) noted that the uniform null distribution is still used most often in research. However, they recommended that researchers should use prior research and theory to select the most appropriate null distribution.

There have been some criticisms of the measure that have lead to subsequent revisions of the computations of r_{wg} and $r_{wg(J)}$. One of the largest issues concerns the appropriateness of $r_{wg(J)}$. Lindell, Brandt, and Whitney (1999) criticized the calculation of $r_{wg(J)}$ for incorrectly applying the Spearman-Brown formula. The Spearman-Brown prophecy is rooted in classical test theory and provides information about the reliability of the scores from a measure. Lindell et al. argued that the Spearman-Brown correction is

not appropriate because r_{wg} and $r_{wg(J)}$ are measures of agreement, not reliability. More recently, however, LeBreton, James, and Lindell (2005), disagreed with Lindell et al.'s critique. Specifically, they demonstrated that the Spearman-Brown prophecy may indeed be appropriate for use with agreement indices. Although there does still seem to be some disagreement about the exact form of $r_{wg(J)}$ to use, LeBreton and Senter (in press) recommended that researchers should use James et al.'s (1993) original formulas for most rating conditions.

Another concern with r_{wg} is regarding the treatment of negative values. Negative values occur when interrater agreement is less than what would be expected due to chance. James et al. (1984) suggested that negative values of r_{wg} should be set to zero. However, Lindell et al. (1999) argued that just as agreement greater than what is expected by chance is captured in the values of r_{wg}, so should values that are lower than what is expected by chance. They suggested that researchers should keep negative values of r_{wg}. As such they advanced the r^*_{wg} index, which is computationally equivalent to r_{wg} for positive values, but allows negative values to remain negative (LeBreton & Senter, in press).

There are other controversies concerning the use of r_{wg} that do not have straightforward solutions. First, it is unlikely that a rectangular distribution will be appropriate; however, correctly specifying the appropriate null distribution is also problematic (e.g., Brown & Hauenstein, 2005; LeBreton & Senter, in press), which is ironic because this feature is the main advancement of r_{wg} over the Finn index. Choosing the incorrect distribution will either under or overestimate values of agreement. For

instance, if a researcher uses a moderately skewed null distribution when in actuality ratings were only slightly skewed, the resulting estimates of agreement will be overstated.

Another problem with r_{wg} is that the observed variance is related to the scale mean in the form of a curvilinear relationship (Brown & Hauenstein, 2005). Specifically, the potential variance of responses is the highest at the middlemost response option. Moreover, the potential variance decreases at both scale endpoints because of ceiling and floor effects. The implication of this relationship is that it is impossible to disentangle actual agreement from the function of the mean rating. Consequently values of agreement are underestimated for the scale midpoint and are overestimated at the low and high ends of the response scale.

In addition, the recommended number of raters needed to provide stable agreement estimates with r_{wg} is unlikely to be attained in most MSF rating scenarios. Both James at el. (1984) and Lindell et al. (1999) recommend having at least 10 raters, although this rule of thumb is often violated (Brown & Hauenstein, 2005). Violating the recommended rater group size can result in attenuated calculations of agreement.

As a consequence of some of the shortcomings associated with r_{wg}, the average deviation index was created (Burke et al., 1999). This index also assesses agreement among multiple raters for their ratings of a single ratee on an interval scale. It is calculated as:

$$AD_{M(j)} = \frac{\sum_{k=1}^{N} \left| X_{jk} - \overline{X_j} \right|}{N} \tag{13}$$

where N is the number of raters for an item j, X_{jk} is the kth rater's rating on item j, and $\overline{X}_j$ is the average of the raters' ratings for the item j. Thus, this index is computed by

summing the deviation scores for *k* raters and then dividing this sum by the number of raters in order to arrive at the average deviation score for a single ratee on item *j*. Burke and colleagues also proposed a parallel measure that uses the raters' median rating rather than the mean rating. The resulting *AD* coefficient is in the units (e.g., five-point or seven-point) in which a ratee was rated (Burke & Dunlap, 2002; Dunlap et al., 2003). Similar to SD_X, *AD* is a measure of dispersion, where larger values correspond to greater dispersion. Also, similar to SD_X, the interpretation of results can be difficult because values of *AD* are not confined to a zero to one scale because the range varies according to the number of response options (Brown & Hauenstein, 2005).

More recently, Brown and Hauenstein (2005) proposed an alternative index of agreement called a_{wg} that is meant for use with multiple raters rating a single ratee on an interval scale of measurement. This index was developed in part because of some of the shortcomings of r_{wg}. The index a_{wg} was derived from Cohen's (1960) kappa, which was originally used to determine the agreement of two raters in assigning a ratee to a categorical condition. Cohen's kappa is calculated as:

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{14}$$

where p_o is the proportion of times that two raters agree and p_c is the proportion of times that agreement would be expected due to chance, which is calculated by summing the proportions assigned to each category. Because the kappa index cannot be appropriately applied to continuous data, Brown and Hauenstein (2005) provided an adaptation of the index that is appropriate for continuous data for assessing rater agreement on a single ratee. Similar to kappa, this metric expresses agreement as a proportion of agreement over the maximum disagreement possible at a given observed mean rating. However, the

a_{wg} statistic adjusts the possible level of disagreement for a continuous scale. The calculation of the statistic is as follows:

$$a_{WG} = 1 - \frac{2 * S_x^2}{\left[(H + L) * M - (M^2) - (H * L)\right] * \left[k/(k-1)\right]} \quad (15)$$

where k is the number of raters, M is the mean rating across raters for a single rater on item x, H is the maximum possible value of the scale, L is the lowest possible value of the scale, and S^2_x is the observed variance of the ratings across a group of raters for item X. Values of a_{wg} range from -1.0 to 1.0, with larger values indicating higher levels of agreement. However, when the mean rating is equal to the upper or lower boundary, the denominator of the equation will be zero; in such cases, agreement should be set to 1.0 (perfect agreement). To illustrate the calculation of a_{wg}, consider a group of four raters who made ratings on a scale that ranged between 1 and 5. The group had an observed variance of .33 and a mean rating of 3.5. In this rating scenario a_{wg} would equal .87:

$$a_{WG} = 1 - \frac{2 * .33}{\left[(5+1) * 3.5 - (3.5^2) - (5 * 1)\right] * \left[4/(4-1)\right]} = 1 - \frac{.66}{\left[(6) * 3.5 - (12.25) - (5)\right] * \left[4/3\right]}$$

$$= 1 - \frac{.66}{\left[21 - 12.25 - 5\right] * \left[1.33\right]} = 1 - \frac{.66}{\left[3.75\right] * \left[1.33\right]} = 1 - \frac{.66}{4.99} = 1 - .13 = .87.$$

Brown and Hauenstein (2005) provided standards for interpreting values of a_{wg} based on prior agreement research (e.g., Kozlowski & Hattrup, 1992; LeBreton et al., 2003). Specifically, they suggested that values between .60 and .69 as should be characterized as weak agreement, between .70 and .79 as moderate agreement, and values exceeding .80 as strong agreement. Values less than .60 should be interpreted as unacceptable levels of agreement; Brown and Hauenstein advised against data aggregation in such cases.

Values of a_{wg} however, can only be interpreted properly for group means that fall within a specified range. Specifically, for smaller rating groups it is impossible to compute agreement for an extreme mean rating that includes both the upper and lower limits of the response choices. Brown and Hauenstein provided the following example, "consider a 5-point scale used by 10 raters. No set of ratings with a mean of 1.3 can include a single rating of 5, the maximum rating of the scale (2005, p.174)." They therefore recommended that the minimum number of raters that researchers should use is roughly one less than the number of response options (i.e., no less than four raters for a five-point scale).

Even when using the specified number of raters, it is possible that a group mean may fall outside of the interpretable range. Brown and Hauenstein (2005) stated that a_{wg} cannot be calculated in these instances. The upper and lower boundaries of useable group means can be calculated with the following two equations:

$$\text{lower boundary} = \frac{L(k-1)+H}{k} \tag{16}$$

$$\text{upper boundary} = \frac{H(k-1)+L}{k} \tag{17}$$

For equations 16 and 17, k is the group size, L is the lowest scale value, and H is the highest scale value. To illustrate, the interpretable range of means for an instrument with five response options will be between 2.3 and 3.6 for groups of three, 2.0 and 4.0 for groups of four, and 1.8 and 4.2 for groups of five. One implication is that for groups where the mean values fall outside of the specified range, researchers cannot calculate a_{wg}. Although, it is reasonable to assume that such groups have high agreement, a quantification of such agreement is not recommended by Brown and Hauenstein.

To calculate agreement across a series of items, Brown and Hauenstein (2005) proposed $a_{wg(J)}$. This multi-item version of a_{wg} is calculated by taking the average of the individual a_{wg} coefficients across items. Specifically,

$$a_{wg(J)} = \frac{\sum a_{wg(1)}}{J} \tag{18}$$

where J is the number of items and $a_{wg(1)}$ is the estimate of agreement for a single ratee on item x. Individual estimates of a_{wg} that are outside of the useable range are treated as missing data. Thus, the level of agreement for a scale of items is the average of interpretable a_{wg} coefficients.

The a_{wg} index has some advantages over r_{wg}. The a_{wg} metric adjusts according to the number of response options such that agreement reflects the possible levels of variance at different scale points. For instance, a mean rating at the middle of a response scale is likely to have greater variance than a mean rating at either extreme of the rating scale. This feature is an advantage over r_{wg} where mean ratings correlate to values of r_{wg}. Brown and Hauenstein (2005) demonstrated that a_{wg} does not have this problem by examining the ratings of experts in the relevancy of situations for a situational judgment test. They found that although values of r_{wg} strongly correlated to the mean rating (r = .63), values of a_{wg} did not (r = -.03). This finding indicates that values of a_{wg} are not influenced by the location of the mean to the extent that r_{wg} is.

However, there are some concerns with a_{wg} that are worth noting. As I discussed earlier, a_{wg} coefficients can only be calculated for group means that are in the interpretable range, which is related to rater group size. In instances where there are relatively small groups, agreement cannot be calculated for extreme ratings. Thus, it may

not be possible to calculate agreement for a large subset of ratings, particularly when extreme ratings are present. Moreover, levels of agreement can be influenced by the number of raters (Roberson, Sturman, & Simons, 2007)

Sampling error and issues of range restriction are also a concern. Sampling error impacts two of the components of a_{wg} (observed mean and observed variance), whereas this is only a concern for the observed variance term when calculating r_{wg} (Brown & Hauenstein, 2005). In addition, the issue of range restriction is also a problem with a_{wg}. Unlike r_{wg}, the impact of response biases such as leniency effects cannot be diminished through the use of a null distribution. One implication is that the level of agreement for the a_{wg} index may be overstated in conditions of range restriction. On the other hand, selecting the correct null distribution for r_{wg} is controversial and can be problematic (Brown & Hauenstein, 2005). Another potential problem with a_{wg} is that values cannot be calculated when group means approach either the upper or lower boundary and are subsequently outside of the interpretable range. This problem is especially relevant for small rater groups that have narrower interpretable ranges.

Comparison of methods. Above I outlined some of the advantages and disadvantages of various estimates of interrater agreement. In addition to those that I noted, restriction of range is also an issue with all measures of agreement. Earlier, I criticized reliability measures for being downwardly biased for situations with range restriction. However, all of the above agreement measures can also be impacted by restriction of range. When any group of raters agrees on a ratee, the range is necessarily restricted. In addition, range may also be restricted in the situations of response biases or leniency. However, unlike measures of reliability, measures of agreement will be

upwardly biased in situations of range restriction. To illustrate, consider the formulas for r_{wg} and a_{wg} (formulas 10 and 15, respectively) which, in their general forms, are calculated as: 1 – (observed variance/expected variance). If ratings within rating groups are relatively homogeneous as a result of leniency or another type of response bias, the resulting observed variance for both r_{wg} and a_{wg} will be low. Consequently, the ratio between the observed and expected variance decreases, and the values of a_{wg} and r_{wg} increases. Thus, the resulting estimates of agreement may be overstated if the observed variance is inflated due to response bias rather than actual agreement. More generally, range restriction will impact any method of assessing convergence. For this reason, depending on the method used, researchers should consider the specific impact of range restriction on their estimates of agreement or reliability.

Bearing these considerations in mind, it was important to choose an index of agreement that was most appropriate for the present research. The purpose of this research was to predict agreement among raters and therefore, agreement was the dependent variable. Some research suggests that alternative methods of assessing agreement yield similar estimates (Roberson et al., 2007). A Monte Carlo simulation that compared estimates of agreement (including SD_x, AD, r_{wg}, r^*_{wg} and a_{wg}) yielded highly consistent results among indices (Roberson et al., 2007). For example, a_{wg} was highly consistent with r_{wg} with a mean correlation of .96.

Despite this finding, there are also some differences between these indices that make a_{wg} more suitable for the current research. One advantage of a_{wg} is that fewer raters are needed to calculate a_{wg} than r_{wg} (4 versus 10, respectively; Brown & Hauenstein; Lindell et al., 1999). Although a_{wg} is unable to account for response biases as r_{wg} can, the

process of selecting the correct null distribution for calculations of r_{wg} is often difficult (Brown & Hauenstein, 2005). Most important, a_{wg} adjusts the calculation of variance in relation to the location of the group mean, and thus is less likely than r_{wg} to be affected by the mean MSF rating (Brown & Hauenstein, 2005). This difference is especially critical because r_{wg} is likely to violate the assumption of homoscedasticity needed for regression because the residuals are likely to vary according to the group's mean rating. Although the violation of homoscedasticity might not be as critical when using agreement to justify levels of consensus (e.g., Roberson et al., 2007), it is particularly relevant in the present research where agreement is being used as a dependent variable. Thus, a_{wg} appears to be the best measure of agreement for the present research based on the options currently available.

Level of agreement in MSF research. Although a_{wg} will be used to assess within-source agreement for the present study, no prior study has used this index with MSF ratings to do so. For this reason, I will review in this section the findings of studies that used other metrics of agreement to examine within-source agreement of MSF ratings. Because agreement indices tend to yield highly similar results (Roberson, Sturman, & Simons, 2006), the few studies that have assessed the level of agreement within rating sources are relevant to the present research.

Fleenor et al. (1996) compared measures of interrater reliability (i.e., Pearson) and interrater agreement (i.e., ICC and T index) of supervisor and subordinates ratings within a single health-care organization. They found low levels of reliability within the subordinate groups (r = .20). In addition, they found low levels of agreement within the subordinate groups using ICCs (ICC(1, 3) = .43). However, their estimates of agreement

were higher using the T index. To assess agreement with the T index, they analyzed their data using a ½-point criterion and a 1-point criterion; these criteria specified the extent to which a rater's average rating on a dimension could deviate from that of another rater and still be considered in agreement. They found relatively high agreement within subordinate's ratings; $T = .63$ for the ½ point criterion and $T = .84$ for the one-point criterion (although they admitted that the one-point criterion was perhaps too generous). The largest criticism of their study, and the use of the T-index, is that agreement is characterized dichotomously.

The remaining two studies that have examined within-source agreement both used r_{wg}. Johnson and Ferstl (1999) calculated r_{wg} using a slight skew null distribution ($\sigma_E^2 = 1.33$) for groups of subordinate raters in a single accounting firm for two different years. The skewed null distribution represents a rating scenario where ratings are negatively skewed. They found that the average agreement was .52 with a standard deviation of .16. However, LeBreton et al.'s (2003) study is perhaps the most comprehensive examination of both reliability and agreement of MSF ratings within and between groups. The purpose of LeBreton et al.'s (2003) study was to challenge the assumption that individuals within rating groups are exposed to similar samples of ratee behavior, and thus, provide more similar ratings than those from other rating groups (Borman, 1997; Murphy & Cleveland, 1995; Tsui & Ohlott, 1988). Instead, they argued that prior findings supportive of this view resulted from restriction of variance and not from differences between rating groups. Restriction of variance in MSF ratings may result both from organizational processes (e.g., selection, counseling, socialization) and rating biases (e.g., halo, leniency), resulting in the attenuation of reliability estimates. Consequently, the low

estimates of within-source reliability may be an artifact of range restriction in ratings and not actual dissimilarity within rating groups.

To test their restriction of range hypothesis, LeBreton et al. (2003) compared Pearson correlations and ICCs, which are both downwardly affected by restriction of variance, with r_{wg}, an index of agreement. They reported results that supported their hypotheses based on a Monte Carlo simulation and a study of graduate students' ratings on the Leader Behavior Questionnaire. However, their study of MSF ratings based on a large sample of managers spanning organizations and industries is most relevant to the present research, and therefore I will discuss it in detail.

In this study, LeBreton et al. (2003) expected that the consistency measures would indicate a low level of rating reliability both within and between rating groups whereas r_{wg} would show high levels of agreement for both. The findings of the study generally supported their restriction of variance hypothesis. Pearson correlations and ICC(1,1) were in the low .30s within peer and subordinate rating groups, suggesting low consistency or reliability. For their measure of agreement, however, they found moderate to high levels of agreement. To assess agreement they calculated r_{wg} using three types of null distributions to approximate agreement for varying levels of response bias: (1) rectangular distribution ($\sigma_E^2 = 2.00$), (2) slight skew ($\sigma_E^2 = 1.33$), and (3) moderate skew ($\sigma_E^2 = 0.90$). The rectangular distribution assumes that raters are equally likely to choose any of the five response options. In contrast, the skewed analyses calculate agreement by taking into account the fact that performance ratings are often negatively skewed because raters are more apt to endorse positive response options.

LeBreton et al. (2003) indicated that the moderate skew scenario is the most probable case for performance ratings. For this reason, I will only report the results of the study that are based on the moderate skew condition. Across the 16 dimensions of managerial effectiveness, they found that the mean within-source agreement ranged between .52 and .74 (M = .60) for subordinates and between .55 and .76, (M = .64*)* for peers. Average agreement between subordinate and peer rating groups was somewhat lower than the within group analyses; agreement ranged between .48 and .73 (M = .57) for different rating dimensions, although whether this difference represents a significant difference from within-source agreement was not reported. Although mean values of agreement did exceed .70 for some rating dimensions, overall agreement did not. Also, although LeBreton et al. interpreted these findings as evidence of high within-source agreement, others might disagree. Specifically, Brown and Hauenstein (2005) recommended that researchers should classify values of agreement between .60 and .69 as weak agreement, and those between .70 and .79 as moderate agreement. Moreover, as I discussed earlier, the r_{wg} index may overestimate agreement in situations of response bias or when the expected variance is incorrectly specified. Thus, researchers should be cautioned against directly comparing the magnitudes of different convergence indices without specifically addressing how the indices are impacted by the rating scenario.

Aside from the magnitude of agreement, it is also possible that the agreement among groups was variable. For peers, the average standard deviation of r_{wg} was .16 for the uniform distribution, .23 for the slight skew, and .28 for the moderate skew. For subordinates, the average standard deviation of r_{wg} was .18 for the uniform distribution, .25 for the slight skew, and .30 for the moderate skew (J. M. LeBreton, personal

communication, May 16, 2006). These findings suggest that significant variability exists within both peer and subordinate rater groups and further indicate the importance of examining this variability in the present study.

LeBreton et al.'s (2003) study was an important development in methods of studying agreement in MSF ratings. The researchers were mainly interested in comparing the average level of agreement with that of reliability as a way to challenge the underlying assumption of MSF ratings that different rater groups provide unique perspectives. In addition, they demonstrated why methods of reliability are inappropriate for examining within-source similarity. However, LeBreton et al. (2003) did not address the extent that there was variability in the level of agreement among rater groups. Yet examining the variability in within-source agreement is particularly important given that other prior research has found a high degree of idiosyncratic rating tendencies (e.g., Greguras & Robie, 1998; Mount et al., 1998; Scullen et al., 2000) which would indicate some between group variability in within-source agreement.

Although LeBreton et al.'s (2003) study was a good first step in exploring the level of within-source agreement, additional research is necessary to better understand agreement, including the extent that the variability surrounding this agreement can be predicted. For this reason, one goal of the present research was to better understand the underlying assumption that ratings from the same MSF source are similar (Borman, 1997). Specifically, to address this issue, I examined possible predictors of within-source agreement, including whether characteristics of the rating group, ratee, and rating dimension were associated with within-source agreement.

Variability in within-source agreement

Most research reviewed to this point concerns the extent to which there is similarity (using either reliability or agreement) within rating groups. Knowing the degree of within-source similarity in ratings provides some valuable information and can be used to help justify the aggregation of a group's ratings. However, this type of research does not examine the extent that within-source agreement varies across rater groups. Yet, there does seem to be some variability in the extent that raters agree with one another, or that some groups rate a focal manager similarly whereas others do not (Greguras & Robie, 1998). Thus, the present research is aimed at examining the predictors that are likely to relate to within-source agreement for different rater groups.

One way to think about the variability of agreement across rating groups is in terms of composition models, or how the relationships between two levels of analysis are specified (Chan, 1998). In the case of MSF ratings, individual ratings, which are at the individual level of analysis, are combined to form an averaged rating that is at the group level of analysis. Most MSF researchers have implicitly proposed what is called an additive composition model whereby individual ratings are combined to form an averaged source rating, often regardless of whether adequate convergence is found. An important consequence of researchers' use of additive models to study MSF ratings is that the variance in levels of agreement, which is considered a by-product of error, has been left unexamined.

Dispersion models, another form of composition models, specifically focus on the dispersion or agreement of responses. In contrast to additive models which treat variance as measurement error, dispersion models specifically examine this variance as the primary construct of interest (Chan, 1998; Feinberg et al., 2005; Klein et al., 2001).

Specifically, dispersion is a group-level variable that describes the variance of individuals' responses within the group. Therefore, research based on dispersion models specifically tests whether the dispersion in ratings is a product of random error or a characteristic of the rating environment.

There are few examples of prior research that have specifically adopted a dispersion model. One such study by Klein et al. (2001) examined the predictors of agreement about the work environment in a large sample of manufacturing plants. Although the study of climate perceptions differs from MSF ratings, the design of the study is applicable. Klein et al. (2001) tested whether demographic homogeneity, work interdependence, social interaction, and survey wording were related to work group members' agreement about their work environment. To operationalize agreement, they computed the group's standard deviation on each item and then averaged the standard deviations for each scale. They also computed $r_{wg(j)}$ as a comparison and found that the group's standard deviations correlated between .8 and .9 with values of $r_{wg(j)}$, with the exception of values of $r_{wg(j)}$ that were outside of the range of 0 and 1.0, suggesting that the two measures offered similar results (note that they did not report how often the scores fell outside of 0 to 1 range, although LeBreton et al. (2003) suggested that this is a rare occurrence). Klein et al. also found that perceptions of the work environment were not related to the demographic characteristics; however, work interdependence, social interaction and survey wording were related to agreement. Thus, by testing the hypothesized predictors of the group's agreement on workplace climate, Klein et al. used a dispersion model to learn more about the predictors of work climate perceptions.

In a more relevant study using a dispersion model, Feinberg et al. (2005) examined how the agreement among subordinates' leadership ratings related to their attributions of transformational leadership. The researchers argued that one aspect of transformational leadership is the ability for leaders to create consensus among their followers, and for this reason, studying the consensus of followers could provide a better understanding of how leadership perceptions are formed. They found that within-group agreement related to their predictors. First, they found a positive relationship between subordinates' agreement on ratings of their supervisors' leadership behaviors (using r_{wg}) and subordinates' mean level of ratings of leader behaviors (r = .50), suggesting that having consensus about a leader's behavior is positively related to perceptions about a leader's overall level of performance. In addition, they found that subordinates' agreement on leadership behaviors moderated the positive relationship between the average ratings of leader behaviors and focal managers' ratings of their transformational style. This finding suggests that both the leaders' behaviors and the extent that subordinates agree in the perceptions of the leader may be important in understanding attributions of transformational leadership.

Feinberg et al.'s (2005) study used a dispersion model in leadership research. They argued that convergence in transformational leadership ratings represented an important characteristic of a leader. Similarly, within-source agreement may also relate to the focal manager's effectiveness. For this reason, the present research will use a dispersion model to test whether characteristics of the rater group and focal manager predict consensus on MSF ratings. However, some of Feinberg et al.'s findings may be inflated because their assessment of agreement and average leadership ratings were from

the same source. Moreover, the values of r_{wg} have been demonstrated to correlate with mean ratings (Brown & Hauenstein, 2005), and thus, one would expect to find a high level of agreement for high mean ratings or when there was a ceiling effect. Despite these methodological and statistical limitations, the overarching concepts of the study have some implications for the analysis of MSF ratings. Moreover, unlike Feinberg et al.'s study, the present research proposes to use a_{wg}. Brown and Hauenstein demonstrated that the location of the group mean does not relate to values of a_{wg} to the extent that it relates to values of r_{wg}.

The aim of the present research was to add to our understanding of agreement within MSF rating groups through the use of a dispersion model. By using a dispersion model to study MSF, I was able to test whether characteristics, such as those of the ratee and the rater group, related to the variability in within-source agreement. As I have discussed earlier, agreement in MSF ratings is important and has consequences for assessing the quality of the ratings (e.g., Borman, 1997; Schmitt et al., 1986). In the following section, I discuss the possible predictors of within-source agreement.

Predictors of Agreement

As previously argued, the vast majority of prior MSF research has focused rather narrowly on comparing the mean level of reliability or agreement between and within rating groups. Moreover, most prior literature, through the use of additive composition models, has assumed that within-source agreement is uniformly high among rater groups. However, there is reason to believe that there is some variability in the extent that rater groups agree in their ratings (Greguras & Robie, 1998), yet little is known about the conditions that relate to this agreement. The purpose of the present research was to

examine the characteristics of the rating group, the ratee, and the rating dimension as predictors of within-source agreement. To explain how these factors relate to level of agreement, in this section, I integrate theory and research on agreement of performance ratings (e.g., Harris & Schaubroeck, 1988; Judge & Ferris, 1993; Landy & Farr, 1980; Murphy & Cleveland, 1995; Tsui & Ohlott, 1988) with Kenny's (1991) weighted-average model of consensus. Specifically, the application of Kenny's (1991) weighted-average model of consensus and accuracy provides a framework to analyze within-source agreement. This section also focuses on identifying characteristics that are likely to relate to the level of agreement. These predictors are based on Kenny's (1991) weighted-average model of consensus along with prior research on interrater agreement and performance ratings.

Most prior research on the interrater agreement or reliability of MSF ratings has focused on reasons why agreement or reliability tends to be low *between* rating sources (e.g., Borman, 1997; Tsui & Ohlott, 1988), citing reasons such as having a unique relationship with the ratee and different models of managerial effectiveness. Yet the finding that agreement or reliability is often low *within* rating sources has rarely been discussed. More generally, theory and research on performance ratings posit some of the possible processes that impact performance ratings. The primary focus of most of the literature on performance ratings is on the level of the rating given. However, the discussion of the sources of rating errors provides a foundation for understanding factors that may relate to within-source agreement. For this reason, I will discuss aspects of the rating process that are thought to relate to rating errors.

Landy and Farr (1980) advanced a fairly comprehensive model of performance ratings to help identify general factors that impact the rating process. They identified four main categories of factors that can impact performance ratings: (1) vehicle or format of the rating instrument, (2) rating process (e.g., level of rater training and rater anonymity), (3) rating context, and (4) roles (e.g., characteristics of rater and ratee, organizational roles, and quality of rater-ratee relationship).

Although the focus of the present study is on the raters and not the rating instrument, there is some research that suggests that the vehicle, or the MSF instrument, impacts within-source agreement. Prior research has generally found that behavioral rather than general trait items (Kaiser & Craig, 2005) and explicit and objective rather than implicit and subjective standards (Schrader & Steiner, 1996) yield higher levels of agreement. In addition, a recent study using a well-established MSF instrument found higher levels of reliability were associated with single-barreled items rather than multi-barreled items (i.e., items that ask raters to consider multiple ideas or concepts) (Kaiser & Craig, 2005). These findings highlight the importance of using a carefully designed MSF instrument to minimize errors in the rating process that result from problematic item formats.

The rating process is generally constant within a MSF administration, so the process should not explain much variation in agreement within rating groups. For instance, the directions given, the use of the data (e.g., developmental or administrative), and rater training will likely be the same across raters who complete the same MSF instrument. For this reason, the rating context should not be a major factor for predicting

varying levels of agreement among rating groups; however, I will briefly review some of the major findings concerning the rating process.

Various contextual variables such as the purpose of the rating and the type of organization have been shown to impact the qualities of MSF ratings. For example, subordinate ratings used for development were found to be psychometrically superior to those used for administrative purposes, although a similar effect was not found for peer raters, who tended to be more reliable (Greguras et al., 2003). Furthermore, when raters were held accountable for their ratings they tended to be of a higher quality (Curtis, Harvey, & Ravden, 2005). Also, an examination of the MSF ratings in different industries found that raters from public sector organizations were more lenient in their ratings than those from private sectors, and that educational organizations were especially lenient (Brutus, Fleenor, & London, 1998).

Another contextual factor important to the rating process is the motivational component of MSF ratings. Murphy and Cleveland (1995) provided a discussion of how raters often have goals that may differ from the organization's goal of obtaining accurate ratings. Raters' goals may supersede those of the organization, affecting the ratings given. Most of their discussion about goals pertained to supervisory ratings (but not peer or subordinate ratings). For example, they suggested that supervisors are likely to have different motives when rating members of their in- and out-group as well as exceptional, average, and poor performers. However, the motives of peers and subordinates are more relevant to the present study and are likely to differ from the supervisor's motives. For instance, it is possible that some subordinate raters are uncomfortable providing upward feedback, perhaps fearing that their ratings may not be anonymous (Conway et al., 2001).

In this situation, subordinate ratings would be influenced by the goal of maintaining a positive working relationship with the ratee, resulting in inflated performance ratings. Conway et al. also speculated that the motives underlying peer ratings are likely to vary as a function of the organizational culture. If the environment is highly competitive, peers may be tempted to rate the ratee lower as a strategy to boost their own standing in the organization. In contrast, for organizational cultures that value teamwork the goals of peers may be to maintain harmony within the group, which is accomplished through giving favorable ratings. These examples of rating source-specific goals and organization-specific goals would both predict that members within the same source would have similar goals. If goals are indeed similar within rating sources, then the individual goals of raters may inflate the overall agreement within rating sources but should not be a significant factor in predicting why some groups of peers and subordinates agree and others do not.

The roles of the raters and ratees, as defined by Landy and Farr (1980) include a wide range of personal characteristics (e.g., demographic, personality, and psychological), organizational roles (e.g., peer or subordinate), and aspects of the rater-ratee relationship (e.g., level of acquaintance, affect). A vast number of research studies have investigated this class of variables, particularly in examining the demographic characteristics of raters.

Landy and Farr (1980) reviewed a number of studies which indicated that demographic characteristics such as gender, race, and age of the rater and ratee related to the *level* of performance ratings given. They commented, however, that because these findings tended to be inconsistent, it was difficult to draw any sound conclusions

regarding the impact of demographic characteristics on performance ratings. Since then, a number of studies have examined how demographic characteristics relate to performance ratings, also with some conflicting results (e.g., Kraiger & Ford, 1985; Mount, Sytsma, Hazucha, & Holt, 1997; Pulakos, Schmitt, & Chan, 1996; Pulakos, White, Oppler, & Borman, 1989; Waldman & Avolio, 1991). Research that has used a relational perspective, which examines the similarity of supervisor's demographic characteristics in relation to those of their subordinates, often finds that similarity in demographic characteristics results in positive outcomes. For instance, Tsui and O'Reilly (1989) found that supervisors in mixed-gender dyads rated their subordinate's performance lower than those in same gender dyads. Therefore, it seems that the characteristics of raters can, in certain circumstances, impact the level of ratings given, however, less is known about how similarity in these roles impacts rater agreement. For example, do groups of raters tend to have higher agreement if they share more similar demographic characteristics?

Another aspect of rater role, the opportunity to observe ratees, has also been discussed as a possible predictor of the reliability of ratings (Dunnette, 1966; Freeberg, 1969; Landy & Farr, 1980; Rothstein, 1990). The assumption is that raters will have more information on which to base their ratings when they have a longer working relationship with the ratee or they have more relevant information about the ratee. When raters lack sufficient information, they are prone to unreliability and are less likely to be accurate and more likely to use stereotypes (Dunnette, 1966). In general, there has been some support for this relationship. When raters have more relevant contact with the ratee, their ratings tend to be more reliable (Landy & Guion, 1970) and more valid (Freeberg, 1969).

In addition, Rothstein (1990) found that performance ratings become more reliable as the opportunity to observe the ratee increased.

The large body of prior research on performance ratings provides a number of possible factors that may influence MSF ratings. However, what is lacking in this research is a comprehensive understanding of the specific circumstances that relate to agreement within rating sources. An integrated theory that includes a number of possible reasons why two or more raters may or may not agree in their ratings is needed to identify possible predictors of within-source agreement. Kenny's (1991) weighted–average model (WAM) of consensus, which was originally proposed to explain when raters will agree in their personality ratings of a ratee, provides a comprehensive framework that can help explain agreement within MSF rating groups. This model has not been used to predict agreement in the performance rating context. Although rating personality is somewhat different from rating performance, the processes that lead to consensus and accuracy may be similar. For instance, some of the predictors of agreement in Kenny's model are similar to those that have been related to the reliability in performance ratings (e.g., acquaintance with the ratee; Rothstein, 1990).

Another benefit to Kenny's WAM framework is that research on this model has been primarily confined to a laboratory setting (e.g., Chaplin & Panter, 1993; Malloy, Agatstein, Yarlas, & Albright, 1997). This controlled environment allowed researchers to make conclusions regarding not just the consensus of raters, but also their accuracy. Although directly assessing the accuracy of MSF ratings for the present studies will not be possible, both Kenny's theoretical model of consensus and subsequent research using this model may help to distinguish how predictors of within-source agreement may also

indicate accuracy or bias. Therefore, for the present research I will combine prior research regarding interrater agreement and reliability with Kenny's weighted-average model of consensus to develop hypotheses and research questions regarding the predictors of within-source agreement.

Kenny's weighted-average model of consensus. Kenny's (1991) WAM of consensus predicts the extent that two raters will similarly rate a ratee using a mathematical model. This model is based on an earlier model of interpersonal perception advanced by Anderson (1981). Although Kenny specifically discussed the WAM as a model of agreement, in actuality, the model expresses reliability (through the use of correlation coefficients). However, conceptually the model is relevant to agreement. Kenny first discussed how individual raters arrive at a rating of a ratee by combining information about the ratee. Then, his model predicted how the following six parameters, (1) the acquaintance of the rater to the ratee, (2) the overlap of stimulus among raters, (3) the shared meaning systems among raters, (4) the communication among raters, (5) the use of extraneous information or bias when rating the ratee, and (6) the consistency of the ratee's behavior, combine to predict the correlation between the ratings of a ratee.

The first stage of person perception, according to Kenny's (1991) model, is how an individual rater combines and weights the pieces of information that the rater knows about a ratee in order to make a rating. Raters incorporate the information that they have about the ratee to make the rating. This information can be divided into what Kenny calls 'acts', which can be based on the ratee's appearance, verbal, or nonverbal behavior. These acts are then aggregated with one another when making a rating. However, acts that are more representative of a trait are given greater weight when making an

assessment. An example of this process is a rater judging a ratee named Kathy on the trait of conscientiousness. The rater notices that Kathy arrives fifteen minutes late, is neatly dressed, and promptly returns a phone call. These pieces of information or acts can be given scale values for conscientiousness. In this case, being late would have a negative scale value, whereas being neat and returning the phone call quickly would have positive values. In addition, these acts could be weighted if one piece of information is considered to be more or less indicative of conscientiousness. Assessments of ratees may be influenced by others. For instance, a colleague may tell the rater that the reason that Kathy was late was because the elevator was broken, which may ultimately impact the rater's assessment of Kathy. Thus, communication from others, by sharing their impressions or providing additional context, may also impact ratings. Finally, error or bias is likely to be present in ratings. Thus, Kenny states that a rater's unique impression of the ratee, which is not based on the ratee's actual behavior, is also incorporated into the rating.

This perceptual process that determines a rater's impression of a ratee, according to Kenny, can be expressed mathematically. However, Kenny did not include the weighting process in his equation, and thus, each act is weighted equally. Specifically, the impression of a ratee for rater *i*, or I_i, is computed with the following equation:

$$I_i = \left[\left(ks_{i0} + \sum_{j=1}^{n} s_{ij}\right)/(k+n)\right] + aI_2 \qquad (19)$$

where, k is the weight given to the unique impression of ratee, s_{i0} is the unique impression for rater i, n is the number of acts that rater i observes, j is the act, s_{ij} is the scale value given to act j by rater i, a is the extent that rater i communicates or is influenced by an outside source, and I_2 is another person's impression of the ratee. Note

that another person's impression will only have an impact when there is communication among raters.

In addition to positing how individuals arrive at their ratings of a ratee, Kenny (1991) also predicted that the extent that two raters make a similar assessment of a ratee is contingent on six parameters. First, when raters have higher *acquaintance* with a ratee, reliability should be greater. Second, the extent of *overlap,* or the degree that each rater views the ratee performing the same behaviors, will positively relate to consensus. Third, when raters have *shared meaning systems* whereby they interpret and label behavior similarly, consensus will be high. Fourth, the amount of *communication* between raters will increase agreement because they are able to share their impressions of the ratee with one another. Fifth, the *consistency* of the ratee will also increase agreement because it is more likely that raters will base their ratings on similar samples of behaviors when the ratee is more consistent. Finally, the sixth parameter states that when raters use *extraneous information*, or information other than the ratee's actual behavior, consensus will be low. Kenny noted that measuring rater's use of extraneous information is often difficult to disentangle from their meaning systems. The similarity of two raters' ratings can be predicted in two situations. First, equation 20 represents the expected correlation coefficient between two raters who have not communicated ($a = 0$):

$$r = \frac{qn\rho_2(1-\rho_1)+n^2\rho_1\rho_2}{k^2+n(1-\rho_1)+n^2\rho_1} \tag{20}$$

where n is the number of acts of the ratee that are observed by the raters, q is the extent of overlap between the raters in viewing the ratee, which is calculated by computing the proportion of acts in which both raters view the ratee, ρ_2 is the correlation between the two raters' scale values that they give to the same act to represent the extent of their

shared meaning system , ρ_I is the consistency of the ratee, and k is the weight for the unique impression or the extent that raters use extraneous information when rating the ratee.

In addition, equation 21 represents the convergence of two raters who communicate with one another:

$$r' = \frac{r + a^2 r + 2a}{1 + a^2 + 2ar} \tag{21}$$

where r is computed as shown in equation 20, and a is the amount of communication between raters. Thus, agreement is predicted to be higher as the communication between raters increases.

Based on equations 20 and 21, one can predict the extent that two raters will have similar ratings for Kathy, the ratee. The better acquainted the two raters are with Kathy, the higher agreement they will have with one another. Also, if they both view Kathy performing similar acts, they will have higher agreement than if one rater only sees Kathy at cocktail parties and the other knows her only as a work colleague. The more that the two raters discuss Kathy, the higher agreement they will have with one another. If Kathy tends to be an erratic or moody person, agreement among the raters is likely to be lower than if she is even-tempered and always pleasant. Finally, the raters will have higher agreement with one another if they are less apt to use extraneous information about Kathy, such as assuming that she is stubborn because she has red hair.

The example I provided above indicates how Kenny's (1991) six components combine to predict the agreement among raters; however, Kenny discussed a number of assumptions regarding the six parameters. These conditions and the equations are more useful in laboratory settings where each parameter can be manipulated and quantified

(e.g., Malloy et al., 1997). In organizational settings, however, it is more difficult to assess parameters, such as the consistency of the ratee or to know exactly how often two raters communicate information that may impact MSF ratings. Thus, it is the concepts, and not the actual equation that Kenny (1991) posited that will be used to specify possible predictors of within-source agreement. In particular, for the present research, I think that the level of acquaintance, shared meaning systems, communication, and the consistency of the ratee are particularly relevant and these constructs will be discussed in greater detail as they relate to my research hypotheses and prior research.

Taken together, research on the performance rating process and Kenny's (1991) weighted-average model can be combined to help identify possible predictors of within-source agreement. This integration is particularly important because little is known regarding the dispersion of agreement within rating sources and what factors may help to partially explain this dispersion. Researchers, when finding somewhat low within-source agreement, have speculated on reasons for the lack of convergence such as differences in the opportunities to observe target behavior or variation in leader-member exchange relationships (Greguras & Robie, 1998; London & Wohlers, 1991; Scullen et al., 2000). However, no prior research has systematically tested the predictors of within-source agreement. The following section identifies how the parameters in Kenny's (1991) weighted-average model of consensus and prior research on performance rating can be used to predict levels of within-source agreement.

Acquaintance. The level of acquaintance, or how well a rater knows a ratee has been linked to the reliability and accuracy of ratings (e.g., Paulhus & Bruce, 1992; Rothstein, 1990). Individual raters are likely to be more consistent in their ratings when

they have an adequate level of acquaintance with the ratee. Similarly, Kenny (1991), in his model of consensus, stated that convergence among raters will increase with higher levels of acquaintance (which is operationalized as the number of acts that a rater observes) because raters become more reliable in their assessments when they have more knowledge about the ratee. Personality research examining the reliability of raters' ratings of a ratee's personality has found some support for this relationship (e.g., Funder, Kolar, & Blackman, 1995; Paulhus & Bruce, 1992). Funder et al. found that sets of raters who were more acquainted with a ratee had greater consistency in their personality ratings (r = .20) than pairs of strangers (r = .12). Research in organizations also found a link between interrater reliability and acquaintance with a ratee (e.g., Rothstein, 1990). In addition, studies by Landy and Guion (1970) and Freeberg (1969) found more reliable performance ratings when raters had more relevant opportunities to observe the ratee.

Rothstein (1990) examined the relationship between the interrater reliability of supervisor ratings and acquaintance with the focal manager. She found a strong asymptotic relationship between acquaintance, which was measured by the ratee's length of supervisory experience, and the reliability of performance ratings made by two supervisors. The reliability of the ratings increased with higher acquaintance; however, they leveled off after time. This finding suggests that the level of acquaintance that a rating group has with the ratee is likely to impact interrater agreement, particularly during earlier periods of acquaintance. One caveat to note with this study, however, is that Rothstein used managerial experience of the ratee as a proxy for acquaintance with the ratee. It is possible that the acquaintance of the rater to the ratee may have been overestimated, particularly for ratees who reported relatively lengthy managerial

experience and for those in organizations with high turnover. Other indirect measures of acquaintance may be flawed as well. For instance, using a time variable (e.g., knowing a ratee for over a year) to gauge acquaintance may also be defective because it fails to consider the depth and relevancy of interactions that enable an accurate assessment of an individual (Freeberg, 1969; Kingstrom & Mainstone, 1985; Landy & Guion, 1970).

Using various operationalizations of acquaintance, prior research in organizational settings has examined how rater acquaintance impacts the *reliability* of supervisory performance ratings, and generally finds that reliability increases when raters are more acquainted with a ratee (e.g., Rothstein, 1990). In the present research, I operationalized acquaintance as the average of how well each rater reported knowing the focal manager. Based on Kenny's (1991) model and prior research, I would also expect higher levels of within-source *agreement* to be associated with greater levels of mean acquaintance with the ratee.

Hypothesis 1: The mean level of acquaintance within rating groups will be positively related to within-source agreement of peers and subordinates.

Demographic composition. The demographic composition of the rater group may also impact the level of within-source agreement for a few reasons. Demography is defined as "the composition, in terms of basic attributes such as age, sex, education level, length of service or residence, race, and so forth of the social entity under study" (Pfeffer, 1983, p. 303). The following section will discuss demographic variables and how they are typically studied in organizational research. Then, I will elaborate in the following two sections about how the degree of demographic diversity of rater groups may relate to two

distinct predictors in Kenny's (1991) model: shared meaning systems and communication among raters.

To the best of my knowledge, no prior research has examined how demographic variables relate to agreement in MSF ratings. Instead, most studies of MSF have examined whether demographic similarity is associated with self-other agreement (e.g., Brutus, Fleenor, & McCauley, 1999; Ostroff et al., 2004), which is a fundamentally different question. Research on self-other agreement focuses on whether demographic characteristics influence the congruence between a manager's self-rating and that of other constituencies whereas the present research is interested in examining how the demographic composition of the rater group relates to within-source agreement.

Pfeffer's definition of demography emphasizes the importance of the composition of the work context. Studying the gender, race or age of a single employee, otherwise known as the categorical approach (Tsui & Gutek, 1999), does not adequately capture an individual's work experience because it is likely that the demographic characteristics of an individual will interact with those of others. Thus, studying the composition of a work group addresses how an individual's characteristics interact with the characteristics of others to shape organizational experiences and outcomes (Tsui & O'Reilly, 1989). Similarly, Kanter (1977) explained how skewed representations of women in organizations can lead to negative experiences for "token" women. Both of these frameworks discuss ways that the distributions of characteristics can affect outcomes at the individual, group, and organizational level, and are considered compositional or structural perspectives (Tsui & Gutek, 1999). In contrast, Tsui and O'Reilly (1989) introduced the relational approach to studying demography, which focuses on the

interaction between an individual and the group or another individual as a predictor of individual, group, and organizational outcomes.

In studying ratings, both perspectives have been used; compositional models examine how group characteristics impact outcomes, whereas relational approaches usually focus on how the similarity of supervisor-subordinate dyads relates to work outcomes. For the present study, the compositional approach will be used to discuss the processes within groups of raters.

Regardless of whether one takes a relational or compositional approach to studying the impact of demography, the underlying mechanism of both processes is the same. At the core, these perspectives state that similar group members tend to be attracted to one another on the basis of their shared traits. Byrne (1971), explained, with his similarity-attraction paradigm how demographic attributes of a group can impact group processes. He stated that similarity in attitudes and personal characteristics leads to perceived similarity between individuals, which in turn, enhances interpersonal attraction. Whereas research on the similarity-attraction paradigm first operationalized similarity in terms of attitudes (e.g., Byrne, London, & Griffitt, 1968; Byrne, London, & Reeves, 1968), additional support for the similarity-attraction paradigm was later found for a variety of demographic characteristics including gender, race, age, organizational tenure, and educational level within organizational settings (e.g., Jackson et al., 1991; McPherson, Smith-Lovin, & Cook, 2001; Mollica & Treviño, 2003; O'Reilly, Caldwell, & Barnett, 1989). Moreover, in an organizational context, similarity in demographic traits has been associated with a number of positive outcomes, including increased affect (Judge & Ferris, 1993; Tsui & O'Reilly, 1989; Wayne & Liden, 1995), higher

performance ratings (Wayne & Liden, 1995), and increased communication (Zenger & Lawrence, 1989).

Additional theories have attempted to further explain how similarity leads to attraction and these outcomes. Both social identity theory (Tajfel & Turner, 1986) and self-categorization theory (Turner, 1987) state that individuals seek to define themselves and do so through the process of social categorization. People choose to belong to or identify with groups that share similar characteristics with them. People may belong to a number of social categories, but tend to identify with those that are salient in a given situation. This group membership allows an individual to satisfy what Tajfel and Turner believe are basic human motives: to build a positive self-image and to maintain self-esteem. In identifying with a group, an individual is also identifying with the positive characteristics associated with the group, which helps to explain why there is attraction between similar group members (Byrne, 1971) and why people are more likely to form friendships with those who are similar (Mollica & Treviño, 2003). In addition to this preference for similar others, or in-group favoritism, is a tendency to devalue the characteristics of the out-group. For this reason, individuals who differ from their work group or supervisor often have more negative work experiences, including higher turnover (Wagner, Pfeffer, & O'Reilly, 1984), role ambiguity, and role conflict (Tsui & O'Reilly, 1989).

Researchers have not looked at whether the demographic composition of a rating group relates to the variance of performance ratings. Instead, most prior research has examined whether demographic similarity relates to higher mean ratings of performance or effectiveness. These studies take a relational perspective in that the similarities

between the rater and ratee are used to predict the level of ratings, which has been supported in many cases (e.g., Bates, 2002; Kraiger & Ford, 1985; Mount et al., 1997; Tsui & O'Reilly, 1989). This type of research does suggest that demographic characteristics relate to the mean level of the rating given; however, there is reason to believe that these characteristics may also predict the level of agreement within groups. Demographic homogeneity has been linked to shared meaning among group members (Rentsch & Klimoski, 2001) and enhanced communication (Zenger & Lawrence, 1989). Both factors according to Kenny (1991) are predicted to increase the consensus of raters.

Shared meaning systems. Having a shared meaning system is a specific type of cognitive process that considers how raters label a ratee's behavior and may be a predictor of within-source agreement. Kenny (1991) suggests that when raters share meaning systems, they will similarly interpret and label a ratee's behavior, which results in making similar ratings. The concept of having a shared meaning system is similar to some work in the managerial effectiveness literature that suggests that raters will agree when they apply similar criteria and weights when making their ratings (Tsui & Ohlott, 1988). Both of these perspectives consider the cognitive processes that are associated with interpreting a person's behavior when making ratings.

The extent to which raters share a meaning system, according to Kenny (1991) can be assessed in one of two ways. The first is the most direct method, whereby techniques of multidimensional scaling are applied. For example, Chaplin and Panther (1993) gave participants a series of behavioral descriptions of traits like friendliness and tidiness and had them rate the extent to which each behavior was indicative of the trait as well as the difficulty and desirability of the item. From these ratings, the researchers

identified how similar rater's profiles were to others by using two Euclidean distance measures as part of multidimensional scaling procedures. Participants with more similar conceptualizations of friendliness and tidiness had higher levels of consensus in rating popular media figures on those traits. Although some researchers have used similar techniques in balanced human resources scorecards (e.g., Becker, Huselid, & Ulrich, 2001), this onerous method of assessing shared meaning systems may not always be possible in organizations, particularly for MSF administrations that require managers to fill out multiple assessments.

However, Kenny suggested that shared meaning systems may also be found between "friends, married couples, or members of the same culture (p. 161)", which is supported by prior research (e.g., Harrison, Price, & Bell, 1998; Rentsch & Klimoski, 2001; Townsend & Scott, 2001). Shared meaning systems or attitudes among similar individuals may result from the interactions among similar individuals. Or, it is also possible that people who share similar characteristics such as race, gender, age, or culture may also have similar upbringings or life experiences that cause them to interpret their environments similarly (Cox, Lobel, & McLeod, 1991; Townsend & Scott, 2001). In either case, a person's demographic traits are likely a proxy for an underlying set of phenomena or experiences that ultimately shape attitudes or meaning systems (Townsend & Scott, 2001).

A number of prior studies have examined the relationship between demographic traits and work-related attitudes. One is a study by Harrison et al. (1998) that investigated the impact of surface-level (e.g., age, sex, and racial diversity) and deep-level diversity (e.g., job satisfaction, supervisory satisfaction, work satisfaction, and organizational

commitment) on work group cohesiveness over time. They found that as the tenure of employees increased, the influence of surface-level diversity decreased and deep-level diversity increased. This finding suggests that surface-level diversity may have the largest impact in the dynamics of relatively new work groups. Although Harrison et al. did not elaborate on these findings, they also found a relationship between the two types of diversities. Specifically, organizational commitment was negatively related to age diversity (-.24) and racial diversity (-.25), suggesting that surface-level attributes also relate to similarity in attitudes. Therefore, even if time diminished the relative impact of certain demographic traits, the relationships between these traits and attitudes may still remain.

A number of other studies have looked at how different demographic groups differ on various attitudes. Research on racial differences often focuses on differences between African-Americans and Caucasians (e.g., Chan, 1997; Cox et al., 1991; Townsend & Scott, 2001). In one such study, Townsend and Scott (2001) found that African-American employees in a sewing plant had more negative attitudes toward their team and valued achievement less than their Caucasian counterparts. Cox et al. (1991) found evidence that African-American team members were more apt to choose cooperative rewards, unlike white team members who preferred competitive rewards. Other research suggests that African-Americans also have less favorable views toward organizational assessment (Chan, 1997; Schmitt & Ryan, 1997). These findings are most likely the result of shared experiences, such as perceived racism in organizational processes, yet they suggest how certain attitudes may differ among different racial groups.

Investigations of gender differences in attitudes also suggest some differences. For instance, a study of gender differences in implicit theories showed that the women, more than men, preferred leaders who were honest, understanding, sincere, non-manipulative (Epitropaki & Martin, 2004). In addition, another study found that women preferred leaders who were more interpersonally sensitive, whereas men were more favorable in their ratings of competitive or aggressive leaders (Deal & Stevenson, 1998). Also, Lefkowitz (1994) found gender differences in job dispositions and attitudes. Specifically, men reported having greater job autonomy, powers and skill variety in their jobs than women. However, after controlling for factors such as job level, the gender differences became non-significant, suggesting that career attainment rather than gender may ultimately shape job attitudes. This finding is supportive of the proxy argument of demographic traits (e.g., Townsend & Scott, 2001) which suggests that individuals whose demographic characteristics differ from one another may also differ in their attitudes.

Other demographic traits such as age, educational, and organizational level also relate to certain attitudes or values. For instance, a study by Deal (2005) investigated the similarities and differences between generations of managers. She found that different age cohorts had unique preferences for their leaders' characteristics. Although older managers rated credibility and trustworthiness as being important, their younger counterparts were more interested in their leader's ability to provide coaching. Whether these different preferences were the result of employees being in different age cohorts or career stages (which is likely to relate to age) is unknown; however, they do suggest that demographic characteristics relate to beliefs about what characteristics leaders ought to possess. Also, Rentsch and Klimoski (2001) found that teams who were more

homogenous in organizational level and educational experience had more similar schemas of teamwork, although similarity in gender and age did not contribute to these shared schemas. Taken together, research on a variety of job attitudes has found significant differences between individuals with different demographic characteristics. Based on prior research, shared demographic traits should relate to shared attitudes and meaning systems, whereas groups who are demographically diverse are less likely to share these similarities.

Communication. The demographic composition of rater groups may also relate to the extent that group members communicate with one another about the ratee. Kenny (1991) suggested that the degree that two raters communicate about an individual will positively impact agreement. Through communicating with one another, raters are able to share information about the ratee. Thus, communication is a way to fill in another person who may not have viewed the same set of behaviors, which ultimately results in greater interrater agreement regarding the ratee. This hypothesis was supported in a series of three laboratory studies which examined how levels of rater communication, in addition to other variables, impacted the extent that raters agreed on their ratings of ratees' personality traits (Malloy et al., 1997).

Kenny's (1991) model of consensus does not discuss whether certain individuals are more likely to communicate. However, according to the similarity-attraction paradigm similar individuals are more likely to interact and converse with one another than dissimilar individuals (Byrne, 1971). Moreover, homogeneous groups are likely to share a similar language and common references than more heterogeneous groups (Zenger & Lawrence, 1989). A few prior studies have found that similarity does indeed

relate to communication at work. For example, similarity in length of tenure and age predicted the frequency with which engineers communicated technical information with other group members (Zenger & Lawrence, 1989). Greater homogeneity in experience, or functional background within top management teams was also associated with more informal communication among members (Smith, Smith, Olian, & Sims, 1994). As a function of the relationship between demographic similarity and communication, it is possible that rater groups who are similar may have closer relationships and are more apt to discuss the focal manager than more heterogeneous groups. One implication of these processes is that within-source agreement should be higher in homogeneous groups than heterogeneous groups (Kenny, 1991).

Most of the research reviewed above examined how various demographic characteristics are linked to specific attitudes or perceptions. These studies are at the individual level of analysis. The present research, in contrast, was specifically designed to examine how characteristics of the group relate to within-source agreement. If *individuals* from different backgrounds are less likely to share similar perceptions and attitudes than similar individuals, then a logical deduction is that heterogeneous *groups* are also less likely to share perceptions than homogeneous groups. One consequence of this process is that more diverse groups should have lower levels of agreement than less diverse groups. For the present research, the composition or the diversity of the group in terms of gender and ethnicity will be assessed by taking the proportion of subgroup members in the group; this method of assessing heterogeneity is typical in research on demographic composition (e.g., Jackson et al., 1991; Townsend & Scott, 2001). For example, a gender-balanced group would have a proportion of .50 women. In addition,

for the demographic variables of age and education, the coefficient of variation will be calculated. A more complete discussion of this coefficient is included in the Method section; however, this variable is essentially the standard deviation of the group on the variable divided by the group mean; values increase as groups become more diverse. The coefficient of variation has also been used often in demographic composition research (Jackson et al., 1991; O'Reilly et al., 1989; Smith et al., 1994), and is recommended by Allison (1978).

The diversity of groups in terms of demographic characteristics is predicted to negatively relate to within-source agreement. First, compared to the shared backgrounds or shared experiences that similar group members are apt have, those from diverse groups are less likely to share such commonalities (e.g., Cox et al., 1991; Townsend & Scott, 2001). Second, diverse groups may be less likely to communicate with one another as compared to homogeneous groups who share a common set of experiences and language (e.g., Smith et al., 1994). Taken together, the findings of research comparing rater diversity and attitudes, meaning systems, and communication provide reasons to believe that more diverse MSF rating groups should be less apt to agree in their MSF ratings of a focal manager than groups who are less diverse.

Hypothesis 2: The diversity of the rating group in terms of gender, race, education, and age will be negatively related to within-source agreement for peers and subordinates.

Differences between peer and subordinate rating groups

One of the main reasons for gathering feedback from multiple perspectives is that different constituents are thought to have varying viewpoints of the focal manager

(Borman, 1997). It is possible that peers and subordinates may approach the rating process differently and that these different vantage points may relate to the extent that raters agree with one another. For instance, there is some evidence that subordinates make less reliable ratings (Greguras, 1998) and may consider more irrelevant characteristics when making MSF ratings (Bates, 2002).

Rating source and level of agreement. Some researchers have found that peer and subordinate raters provide ratings of varying quality. For instance, prior literature on MSF suggests that the two groups differ in the psychometric qualities of their ratings, with peer ratings being somewhat superior to that of subordinates (e.g., Bates, 2002; Greguras & Robie, 1998). Greguras and Robie (1998), for example, estimated that eight peers and nine subordinates were needed to attain an acceptable level of reliability on a five-item scale. London and Wohlers (1991) also found that the reliability of peers was higher than subordinates (r = .24 and .18, respectively). The same was true for LeBreton et al.'s (2003) study of agreement (peers r_{wg} = .64, subordinates r_{wg} = .60), although they did not test whether the difference in agreement between rating groups was significant. In addition, a study using generalizability theory found that the reliability of subordinate ratings was impacted by the purpose of the ratings whereas the quality of peer ratings was unaffected by rating purpose (Greguras et al., 2003). Specifically, subordinate ratings were more reliable for developmental, rather than administrative purposes. Taken together, these research findings suggest that the quality of peer ratings is somewhat higher than that of subordinate ratings, and subsequently I expect to find higher levels of agreement among peers rather than subordinate groups.

Alternatively, it is possible that the reason that subordinate raters tend to be in less agreement is because they have less training or managerial experience to make psychometrically sound ratings (Bates, 2002). To rule out the possibility that lower agreement is a function of experience, I plan to control for the average tenure of the rater group when performing analyses that compare the WSA of peers and subordinates.

Hypothesis 3a: Within-source agreement will be higher in peer, as compared to subordinate rating groups after controlling for the average tenure of the rater group.

Another possible reason for the lower reliability and agreement within subordinate rating groups, compared to peers, is their use of extraneous information. Recall that Kenny's (1991) model of consensus states that agreement will be lower when raters apply extraneous information to their ratings than when they do not. There is some indication that subordinates may be more influenced by irrelevant factors than peers. For instance, Bates (2002) found that compared to supervisors and peers, subordinates' ratings had stronger relationships with irrelevant factors. Specifically, subordinate MSF ratings were related to their liking and demographic similarity to the ratee, whereas supervisor and peer ratings were not. This finding provides one explanation for why subordinates are more variable in their ratings. If subordinates are more likely to incorporate irrelevant information into their ratings, it makes sense that the relationship between the rater group's demographic traits and within-source agreement is likely to be stronger for subordinates than peers. Specifically, examining the relationships between demographic characteristics and MSF ratings could indicate that subordinate raters are

more influenced by their group's demographic composition (e.g., through group processes such as communication and liking) than are peers.

Hypothesis 3b: Rater-group source will moderate the relationship between demographic characteristics and within-source agreement such that the relationship between demographic diversity and within-source agreement will be stronger for subordinates than peers.

Opportunity to observe behavior. Although I expect peers to have higher levels of agreement than subordinates, the level of agreement within each rating source may relate to the type of behavior being rated. Specifically, the rater group's opportunity to observe particular types of managerial behaviors may influence agreement. If a particular rater group is not able to witness the ratee engaging in a type of behavior, there is apt to be greater error in their ratings, and therefore, less agreement within the group for that particular behavioral domain (Kenny, 1991). Moreover, the opportunity to observe behaviors may be different for peers and subordinates because of their different perspectives (Borman, 1997). Thus, peers and subordinates may differ in their opportunities to observe various managerial behaviors.

Peers are in a unique position to view the focal manager's behavior. For instance, Organ (1997) stated that peers are particularly well-suited to view organizational citizenship behaviors because peers may call on each other for help. In addition, peers, because they often have similar roles and training, are thought to be particularly adept at judging technical competence and separating effort from performance (Fletcher & Baldry, 1999; Klimoski & London, 1974). Thus, it makes sense that peers will have ample information on which to rate an individual in areas such as building relationships,

flexibility, resourcefulness. However, peers will probably have less information about how he or she provides development and interacts with his or her direct reports.

Subordinates are likely to have ample experience viewing the ratee's leadership ability (Conway et al., 2001). I would expect a high level of agreement among subordinates on rating dimensions such as hiring a competent staff. On the other hand, there are other areas where subordinates may not have ample opportunities to observe behavior. For instance, subordinates are not likely to have adequate information about the manager's business acumen, particularly if the ratee has a different skill set from his or her staff.

Although I have advanced some possible domains in which peers and subordinates may have high and low opportunities to observe various types of behaviors, there is not adequate evidence to make specific hypotheses for each rater group's opportunity to observe dimensions typical of MSF instruments. For this reason, the present research gathered data from subject-matter experts about the extent that each rater group is able to observe specific types of managerial behaviors. These average *opportunity-to-observe* ratings were ranked separately for peers and subordinates and then classified as either a *high* or *low opportunity-to-observe* dimension. It was predicted that rater groups would have higher agreement when they have more opportunities to observe the behavioral domain being rated.

Hypothesis 4: Within peer and subordinate groups, agreement will be higher for *high opportunity-to-observe* rating dimensions as opposed to *low opportunity-to-observe* rating dimensions.

Consistency of the focal manager

Kenny (1991) stated that when raters evaluate more consistent ratees they should have higher levels of agreement. In the following sections, I discuss two characteristics that may relate to the consistency of a ratee. First, personality research has discussed the characteristics of people who are more judgable (e.g., Funder et al., 1995). Judgable people tend to provide raters with more information and tend to be more open than less judgable people. However, no prior research has examined the relationship between the focal manager's personality and within-source agreement. Second, the overall performance of the focal manager may relate to the level of agreement among raters (Feinberg et al., 2005). It is possible that very high and very low performing ratees deliver consistent levels of performance, whereas mediocre managers may vary in their results. The variability in rating a mediocre manager may add ambiguity to the rating process, and subsequently may be associated with less agreement among raters.

Personality of the focal manager. Kenny's (1991) WAM of consensus predicts that ratees who are more consistent are easier to rate. The consistency of a ratee is similar to prior personality research on meta-traits (e.g., Baumeister & Tice, 1988) and the judgability of a ratee (e.g., Colvin, 1993; Funder & Colvin, 1991). Specifically, these areas of personality research and theory state that people who have a particular set of characteristics are easier to rate than those who do not. The ease of rating such people is a function of the focal manager giving raters relevant or consistent information. However, another possible relationship between the ratee's personality and rater agreement is that some personality traits have higher agreement associated with them as a function of implicit theories, which associate specific behaviors with the ratings of traits (e.g., Borman, 1987; Lord et al., 1984; Phillips & Lord, 1986). Thus these two explanations

differ. According to Kenny (1991) the judgability of a ratee pertains to the consistency or amount of information that the individual provides to his or her raters. In contrast, proponents of implicit theories alternatively explain that agreement among raters is a function of raters having shared theories about how particular behaviors relate to the rating domain (e.g., Nathan & Alexander, 1983; Phillips & Lord, 1986).

In both cases, the relationship between personality and agreement presupposes that a ratee's personality will impact their behavior to an extent. However, there is some debate regarding the relative influence of situation and personality in predicting an individual's actual behavior (e.g., Kenrick & Funder, 1991; Mischel, 1969, 1977). In particular, research suggests that when there is a strong situation, or one that has highly structured rules of behavior, an individual's personality is less apt to exert influence on behavior (Mischel, 1977). Although some prescribed behaviors are expected within any work environment, the strength of these expectations is likely to vary. For instance, Beaty, Cleveland and Murphy (2001) stated that weak organizational contexts occur when an organization does not adequately develop their employees, supervisors have infrequent interactions with subordinates, or when environmental cues about desired behaviors are sporadic or inconsistent. With larger spans of control, higher turnover, and more telecommuting in organizations today as compared to the past (e.g., Kurland & Bailey, 1999; Sullivan, 1999; Thomas, 1999), it is reasonable to conclude that the situations in many work environments are apt to be relatively weak and that personality will most likely influence people's behavior to an extent.

If personality is likely to impact behavior in work contexts, how might the expression of personality relate to the agreement of MSF raters? Kenny (1991) stated that

a ratee's consistency will positively relate to the agreement of personality ratings. For instance, he theorized that it is easier for raters to judge an individual that is always agreeable, regardless of the situation, than one who is more variable. Personality research has also shown that some people express traits more consistently than do others. Baumeister and Tice (1988) termed the propensity to consistently express a trait as a metatrait. The concept of metatrait is related to the behavioral consistency and judgability of a ratee (Funder, 1995); those who have a metatrait tend to act more consistently across situations, and are therefore more easily judged than those without a metatrait.

This idea of behavioral consistency has some important implications for rater agreement on MSF ratings. Some may argue that MSF ratings, which are usually based on behaviors and managerial skills, are different from rating personality traits. This difference is important to note; however, it is possible that personality traits that impact the consistency of behavior in general may relate to the consistency of managerial behavior as well.

Most prior research on metatraits is not relevant to the present study. Many studies of metatraits were confined to a few behavioral domains that were unrelated to managerial behavior, and findings were at times contradictory (for review, see Chaplin, 1991). However, a more relevant study by Colvin (1993) provided one of the most complete investigations about what types of people tend to be more easily judged.

In this study, Colvin (1993) had participants describe themselves using a 100-item Q-sort that contained personality trait descriptors. Additionally, two well-acquainted peers rated the ratee, as did coders who viewed videotaped interactions of the ratee. In comparing self-peer, peer-peer, and peer-coder ratings of the personality profiles, Colvin

found reliable indications that some individuals were more judgable than others because the three criterion measures correlated significantly. In addition, behavioral descriptions significantly related to ratee judgability. More judgable individuals described themselves as sensitive, warm, compassionate, and socially skilled. In contrast, less judgable ratees endorsed being distrustful, defensive, and having fluctuations in mood. Similar types of descriptions emerged from both peer and coder ratings. In addition, the Big 5 factors also related to judgability. Extraversion, conscientiousness, and agreeableness were positively related to judgability and neuroticism, negatively. Scores on the Hogan Empathy scale and femininity also positively related to judgability, but those on masculinity and self-monitoring did not. Coders, in particular, associated many extraverted types of behaviors with more judgable ratees, including being socially skilled, talkative, high in energy and enthusiasm, and making eye contact, whereas the individuals they judged as being unexpressive, timid, and uninterested had lower levels of judgability. Taken together, these findings provide rich descriptions of how judgable people describe themselves and are described by peers and anonymous raters. Also, I should note that Colvin's research used correlation coefficients rather than a measure of agreement. However, this methodology would have more of an impact on the magnitude of his findings than the specific predictors of judgability.

Funder (1995) integrated findings from Colvin and other research to theorize why some people are better targets than others. He explained that a 'good target' is one who provides raters with behavior that is both informative and relevant. Informative people are active and vibrant and give raters more cues about who they are. Less informative people tend to be passive and quiet. This idea closely relates to Colvin's findings that

judgable individuals have more extraverted behaviors. Aside from the number of cues given, some ratees also give more relevant or truthful cues to others. For instance, individuals who are defensive because they are less well-adjusted (Colvin, 1993) do not provide equally relevant information to raters as opposed to those who are more forthright in their actions. Thus, people who give less relevant behavioral cues are less likely to have raters agree about their personality.

Based on Kenny's (1991) WAM of consensus and the research findings of Colvin (1993) and Funder (1991), there is reason to believe that within-source agreement is likely to relate to the focal manager's personality. This is because there is evidence that links higher expressions of the traits extraversion, agreeableness, and conscientiousness with the judgability of a ratee. The consistent finding that extraversion is related to agreement (Colvin, 1993; Funder, 1995) is particularly relevant to the managerial population. Based on Colvin's research (1993) and Funder's (1995) model, it is likely that more extraverted managers will provide raters with more information about how they perform at work. For instance, extraverts are talkative and are likely to provide more information about the projects that they are working on, including the actions that they are taking and the challenges they are facing. Introverts, in contrast, are less apt to give the rich details surrounding events at work. As a consequence, co-workers are less apt to know exactly how an introvert dealt with a conflict or a difficult situation unless they were able to directly observe the action. This uncertainty associated with rating an introvert adds error to the rating process, which should increase disagreement among raters. Thus, the extent that a focal manager is extraverted is likely to impact the level of agreement within rating sources.

An alternative explanation of why agreement among raters is associated with the ratee's extraversion is related to implicit personality and implicit leadership theories. Prior research on personality ratings has found that rater's implicit theories, or associations of specific behaviors to traits (e.g., talkativeness and extraversion), impacts personality ratings. For instance, Mehl, Gosling and Pennebaker (2006) found that ratings of extraversion correlated with the ratee's amount of talking, laughing, and time alone (negatively), suggesting that raters' implicit theories of extraversion were related to the behaviors of talking, laughing, and level of sociability. Thus, the amount of information that a ratee provides is apt to be confounded with ratings of extraversion.

In addition, Epitropaki and Martin (2004) examined the factor structure of a Offermann, Kennedy, and Wirtz's (1994) ILTs scale. They found that a six-factor structure of ILTs that was generalizable across different employee groups and time periods. Specifically, the factor structure suggested that sensitivity, dedication, intelligence, and dynamism are protypical leader traits and tyranny and masculinity are anti-protypical, or negatively related to leader prototypes. In particular, the dynamism factor includes traits such as energetic, strong, and dynamic. These traits are similar to descriptions of extraversion, and thus, it is certainly reasonable to deduce that extraversion is a component of rater's ILTs. Consequently, raters' agreement for an extraverted individual may be based on their leader prototypes, rather than the ratee's actual behavior.

One implication is that concepts of judgability and implicit theories are confounded when making ratings of a ratee's extraversion. An individual who gives raters more information is also one who fits the prototype of an extraverted manager.

However, there is less evidence to suggest that extraverted behaviors will strongly relate to the implicit theories regarding all managerial competencies such as having business acumen and being self-aware. Put another way, although extraversion has been shown to relate mildly to overall job performance (Barrick & Mount, 1991) and is a component of ILTs (Epitropaki & Martin, 2004), it is less likely that extraverted behaviors will be strongly related to judgments of effectiveness on all facets of managerial performance.

To investigate the possible confound when using extraversion as a predictor of within-source agreement, I make two hypotheses with competing theoretical explanations. First, if the trait of extraversion is confounded with ratings of managerial effectiveness because of the amount of information that extraverts compared to introverts provide to raters, the addition of other personality traits as predictors can help to challenge this explanation. Specifically, Colvin's research suggested that individuals that are agreeable and conscientious also tend to be more judgable because they provide raters with truthful or relevant information. The possible confound between extraverted behaviors and amount of information is not an issue for the traits of agreeableness and conscientiousness. Extraverted individuals are apt to be more informative than their introverted counterparts because of the extent that they talk and provide others with information (Mehl et al., 2006). In contrast, both agreeable and disagreeable individuals provide raters with similar amounts of information (Paunonen, 1989). An agreeable ratee may smile, whereas a disagreeable individual may furrow his or her brows or scowl. Similarly, individuals low and high on conscientiousness also provide similar amounts of information to their raters. Thus, finding a consistent pattern of agreement associated

with extraversion, agreeableness, and conscientiousness, would provide more support for Funder's model of judgability rather than the implicit theories explanation.

Second, I stated above that ratings of managerial performance are different from personality ratings, and that extraversion is not likely to relate to some specific dimensions of managerial behavior. Research on the relationship between the scales of Benchmarks®, a MSF instrument, and the extraversion of focal managers supports this assertion (Center for Creative Leadership, 2004). Although the relationships tended to be small, peer ratings were significantly and positively related to the focal manager's extraversion for the scales of *Participative Management* (r = .16), *Change Management* (r =.12), *Leading Employees* (r = .08), *Confronting Problem Employees* (r = .08), *Doing Whatever it Takes* (r = .06), *Decisiveness* (r = .06), and *Compassion and Sensitivity* (r = .06). These relationships make sense because managers must engage with others to successfully provide colleagues with information, which is an important component of participative management. Involving others in change initiatives is also crucial for responding effectively to change. Also confronting and leading employees require assertive behaviors and are likely to come more naturally to extraverted individuals rather than introverts. On the other hand, peer ratings from the nine remaining scales did not correlate significantly with the extraversion of the ratee: *Resourcefulness, Being a Quick Study, Building and Mending relationships, Balance Between Personal Life and Work, Self-Awareness, Putting People at Ease, Differences Matter*, and *Career Management*. The lack of a relationship between these dimensions and extraversion makes sense. For instance, being resourceful, quickly learning business knowledge, and being sensitive to

diverse backgrounds or colleagues' obligations outside of work should not have much in common with a manager's level of extraversion.

Correlations between the ratee's extraversion and managerial competencies were also computed for subordinates (Center for Creative Leadership, 2004). Subordinate ratings, however, only related significantly with 2 of the 16 Benchmarks® scales. Similar to peers, subordinates rated more extraverted managers as being more effective in *Participative Management* (r = .09). In addition, they perceived more extraverted managers as being less effective on *Being a Quick Study* (r = -.09) (which was unrelated to extraversion for peers). Perhaps introverts are perceived by their subordinates as engaging in more introspective behaviors that allow them to quickly learn business-related knowledge. Ratings from subordinates for the other fourteen scales did not relate to the extraversion of the focal manager.

The relationships between the focal manager's extraversion and MSF ratings do not indicate whether the relationship between ratings of effectiveness and extraversion are the result of raters' implicit theories about what makes managers effective at particular types of behaviors or if extraverted managers are in fact more effective in these areas. However, these findings suggest that if implicit theories do impact ratings, and the level of agreement among raters, they should have the largest impact on the rating dimensions that have shown the strongest relationships with the focal manager's level of extraversion. On the other hand, if a focal manager's extraversion has implications for the consistency and amount of information that he or she provides, in general and regardless of what is being rated, then one can expect extraversion to relate to within-source agreement on all rating dimensions. Moreover, finding similar relationships between

agreement on agreeableness and conscientiousness would further support Kenny's (1991) WAM of consensus. Thus, I advanced the following two hypotheses with competing theoretical explanations.

Hypothesis 5a: The level of the focal manager's extraversion, agreeableness and conscientiousness will be positively related to within-source agreement of peers and subordinates for all rating dimensions.

Hypothesis 5b: The relationship between within-source agreement for peers and subordinates and the focal manager's extraversion will be stronger for rating dimensions that have been shown to relate to the focal manager's extraversion.

No prior research has examined whether rater agreement relates to the focal manager's personality traits. In addition to extraversion, agreeableness, and conscientiousness, it is certainly possible that other personality traits may also relate to the agreement among MSF raters. Some traits may polarize raters such that some feel that a trait is an asset to managerial performance and others feel that it is a drawback. For instance, a manager's need for control may seem positive to some, but other raters may feel the trait is detrimental to performance. Also, ratees who are highly self-aware may have higher levels of agreement associated with them compared to ratees who are less self-aware because they are better able to adjust their interactions to act appropriately regardless of the situation (Fleenor, et al., 1996). Because this area has not been studied, I propose doing some exploratory research to identify other personality traits that relate to rater agreement in MSF ratings.

Research Question 1: Do other personality traits of the focal manager relate to within-source agreement of peers or subordinates?

The level of acquaintance that raters have with the focal manager may also alter the relationship between personality and within-source agreement. Although ratees that possess particular personality traits may differ in the amount of information they provide to raters, the impact of this process is likely to lessen over time. For instance, it seems reasonable to predict that extraverted ratees are apt to provide colleagues with information about who they are quickly. In contrast, introverts may not provide information as quickly, but over time, the amount is likely to approach that of their extraverted counterparts. For this reason, I predict that the relative impact of a ratee's personality in predicting within-source agreement will decline as the level of acquaintance within the rater group increases. This prediction supports Kenny's (1991) model which states that higher agreement should occur when raters have more information on which to base their ratings.

> **Hypothesis 5c:** The strength of the relationship between the focal manager's personality and within-source agreement will be moderated by the extent that peers and subordinates are acquainted with the ratee such that the relationship between personality traits and agreement will be stronger with lower levels of acquaintance.

Performance of the focal manager. The relative effectiveness of a ratee may also relate to the level of agreement within rater groups. Very high and very low performing ratees may be perceived as more consistent in their behavior than mediocre managers. Whereas very high performing focal managers consistently deliver excellent results, and very poor performers deliver consistently poor results, mediocre ratees are likely to have more variable track records. The variance in the effectiveness of these mediocre

individuals adds a level of complexity to the rating process that may ultimately cause greater disagreement among raters (Kenny, 1991).

There is some initial support for this relationship between agreement and effectiveness for ratings of transformational leadership. Feinberg et al. (2005) found a positive relationship between within-group agreement on ratings of supervisors' leadership behaviors and mean ratings of leadership behaviors, suggesting that raters have higher agreement when rating a more transformational leader. However, Feinberg et al.'s study used r_{wg} to calculate agreement, which has been shown to correlate with the group mean (Brown & Hauenstein, 2005). Thus, such a relationship may be inflated based on the scale dependence of r_{wg}. I plan to use a_{wg} to assess agreement, which is impacted less by the location of the group's mean rating. In addition, Feinberg et al.'s research was confounded because the leadership ratings and estimates of agreement were both from the same group of raters. However, I plan to use an external measure of effectiveness from the focal manager's supervisor, which will be independent of peer and subordinate ratings. Note that agreement should increase as managers become progressively better or worse. However, there is likely to be some range restriction in the managerial population (e.g., Mount, 1984) with very few low performing managers in my sample. Therefore, I predict finding a direct rather than curvilinear relationship between agreement and effectiveness.

Hypothesis 6: The within-source agreement of peers and subordinates will be positively related to the focal manager's overall effectiveness.

Relative importance of predictors of within-source agreement

I proposed a number of possible predictors of within-source agreement that relate to the factors from Kenny's WAM of consensus (1991), including the level of acquaintance, shared meaning systems, communication among raters, and the consistency of the focal manager. As the first study to examine the predictors of agreement within groups in this manner, it is important to investigate the relative strength of the predictors discussed. Research on Kenny's model has typically isolated one or two parameters per study (e.g., Malloy et al., 1997). Therefore, there is no direct evidence, to the best of my knowledge, regarding which predictors have stronger relationships with agreement. For this reason, I will report the relative importance of each predictor in predicting within-source agreement using relative weights analysis (e.g., Johnson, 2001).

Because prior research has not specifically examined the predictors of within-source agreement, it is difficult to make predictions about the relative strength of the various predictors discussed. It is possible that there will be relatively small effects for the demographic composition variables because they are proxy measures for shared meaning systems and level of communication, and thus, they do not directly capture these concepts. Moreover, variability in demographic composition may decrease over time (Schneider, 1987; Schneider, Goldstein, & Smith, 1995). However, the variables that are associated with the focal manager's personality and overall performance are direct measures, and therefore may be more robust predicators of agreement. In addition, because the present study used a direct assessment of the quality of the acquaintance, rather than a specified length of time knowing the focal manager, stronger relationships between acquaintance and agreement may be found than in prior studies (e.g., Rothstein, 1990).

Research Question 2: What will be the relative importance of the predictor variables (i.e., acquaintance, demographic diversity, focal manager personality, focal manager performance) in predicting within-source agreement of peer and subordinate groups?

CHAPTER 3: Method

Sample and Procedures

Data were collected from 33,696 focal managers from private organizations located in the United States. The managers completed the Center for Creative Leadership's (CCL) multisource feedback instrument, Benchmarks® between the years of 2000 and 2007. In addition to the managers' self-ratings, developmental feedback ratings were provided by their supervisors, peers, and subordinates. A subset of 7,257 focal managers provided personality assessments. Because the purpose of this research was to predict the variability of agreement, it was essential to use MSF data that minimized the leniency biases that sometimes occur with MSF ratings. For this reason, I chose to restrict my sample to focal managers from the private-sector because prior research suggests that MSF ratings from private organizations are less lenient than those from public organizations (Brutus et al., 1998).

Demographic characteristics of the focal managers in my sample are reported in Table 1. Of the 33,696 focal managers in my sample, 68% were male and 82% were Caucasian. Focal managers were on average 42 years old. Participants spanned organizational levels; 4% were from top management, followed by 25% from the executive level, 43% from upper-middle management, 36% from middle management, and 2% from first-level management. Most reported having moderate levels of experience in their job (56%); 20% reported being very experienced and 24% inexperienced. The focal managers worked in diverse private industries, including manufacturing (29%), finance/insurance/banking (23%), health (8%), transportation/communication/utilities (8%), and wholesale/retail trade (7%).

Raters reported how well they knew the focal manager on a four-point scale ranging from 1 (I hardly know this person) to 4 (I know this person extremely well). Of raters in my sample, most (70% of peers and 69% of subordinates) rated their acquaintance as a 3, indicating that they knew the focal manager moderately well. Peer groups ranged in size from 1 to 18, with a median of 4 raters (M = 3.75). Subordinate groups were similar in terms of size, with a range of 1 to 25 raters and a median of 4 raters (M = 3.83).

Pilot Study. Hypothesis 4 stated that rater groups would have higher agreement when they have more opportunities to observe the behavior being rated. Moreover, it is likely that peers and subordinates differ in the extent that they view a focal manager performing particular types of behaviors. To determine the extent that peers and subordinates observe each Benchmarks® dimension, I administered a questionnaire to 10 advanced Baruch College Industrial/Organizational Psychology doctoral students with organizational work experience. The raters received a sample of 40 Benchmarks® items representative of the 16 Benchmarks® Section 1 scales. I included at least two items per scale in the questionnaire to achieve reliable estimates of opportunity to observe for each dimension. Because the questionnaire used copyrighted items from Benchmarks®, Appendix A is a shortened version of the actual questionnaire.

The students rated how frequently peers and subordinates view specific managerial behaviors on a five-point scale (1 = *very infrequently,* 5 = *very frequently*). I averaged their ratings for each dimension to classify the dimension as either a *low* or *high opportunity-to-observe dimension*. One rater failed to complete the survey; subsequently there are 10 ratings for peer behaviors questions and 9 for subordinates. I calculated two-

way random effects intraclass coefficients (ICCs), using the agreement standard, to assess the average agreement of the students' ratings for each dimension (McGraw & Wong, 1996). The raters failed to agree at an acceptable level for three rating dimensions for the peer and subordinate perspectives. From the peer perspective, the dimensions with low agreement were *Straightforwardness and Composure*, *Putting People at Ease*, and *Differences Matter* (i.e., ICC(2, A) values ranged between 0 and -2.16). Similarly, raters did not agree about how often subordinates observe their supervisors performing behaviors from the *Building/Mending Relationships*, *Straightforwardness and Composure, Balance Between Personal and Work Life* dimensions; values of ICC(2, A) ranged between .05 and -.05. The remaining dimensions had acceptable levels of agreement (Peers: M = .51, SD = .16; Subordinates: M = .64, SD = .19) and were used to classify rating dimensions as *high* and *low opportunity-to-observe dimensions*.

Measures

Managerial behaviors. Benchmarks® is a multisource feedback instrument that was designed to assess competencies that are related to managerial development (Center for Creative Leadership, 2004; Lombardo, McCauley, McDonald, & Leslie, 2001; McCauley & Lombardo, 1990; McCauley, Lombardo, & Usher, 1989; Zedeck, 1995). The content of Benchmarks® was based on interviews with executives who discussed pivotal experiences in their careers and what they learned from these experiences (Lindsey, Homes, & McCall, 1987). Based on content and factor analysis of these interviews, researchers at CCL developed Benchmarks®, which is divided into three main sections. I used *Section 1* and *Section 3* for the present research. *Section 1: Leadership Skills and Perspectives* contains 16 dimensions that relate to critical

developmental experiences. *Section 3* asks raters to assess the focal manager on eight overall effectiveness items. These ratings are for research purposes only and are not used during the feedback process. Benchmarks® was revised in 2001 to ensure that the items were not racially biased (Lombardo et al., 2001) as the original instrument was largely based on the experiences of white males (Lombardo et al., 2001; Zedeck, 1995).

There is evidence that the ratings from Benchmarks® are psychometrically sound (Center for Creative Leadership, 2004; Lombardo et al., 2001; McCauley & Lombardo, 1990). The Section 1 scale scores have high average internal consistency, with alphas ranging between .79 and .93 (Center for Creative Leadership, 2004). Benchmarks® ratings have also been shown to relate to a number of career outcomes. Supervisor ratings of the focal manager's ability to be promoted one-level, overall ability to be promoted, and long-term professional success significantly correlated with all of the 16 dimensions (Center for Creative Leadership, 2004). Ratings of current performance also related to all Benchmarks® dimensions with the exception of *Balance Between Personal Life and Work* (Center for Creative Leadership, 2004; McCauley & Lombardo, 1990).

The 16 leadership competencies that comprise Section 1 include: *Resourcefulness, Doing Whatever it Takes, Being a Quick Study, Decisiveness, Leading Employees, Confronting Problem Employees, Participative Management, Change Management, Building/Mending Relationships, Compassion and Sensitivity, Straightforwardness and Composure, Balance Between Personal and Work Life, Self-Awareness, Putting People at Ease, Differences Matter* and *Career Management*. In total, there are 115 items in Section 1, with each dimension containing between 4 and 14 items. For each item, focal managers are rated by peers and subordinates on the extent that they engage in a behavior

ranging on a scale of 1 (*not at all*) to 5 (*to a very great extent*). Sample content for Section 1 items is presented in Appendix B.

For the present research, Benchmarks® ratings from each rater were averaged within each rating dimension. For instance, the four items that form the *Decisiveness* dimension were averaged for each rater. These dimension scores were used to compute agreement within rating groups. This treatment is identical to LeBreton et al.'s (2003) study, which also used Benchmarks® data to calculate within-source agreement using r_{wg}.

Overall managerial effectiveness. Consistent with Graves, Ohlott, and Ruderman (2007), I measured the focal manager's overall effectiveness with three items from Section 3 of Benchmarks®. These items, which were rated by the focal manager's supervisor, included the focal manager's performance as a leader (1 = *among the worst*, 5 = *among the best*), performance in his or her job (1 = *among the worst*, 5 = *among the best*), and the likelihood that the focal manager's career will be derailed (1 = *not at all likely*, 5 = *almost certain*; reverse scored). Ratings for these three items were internally consistent in the present study ($\alpha = .81$), although prior research using the same items found higher levels of internal consistency (Graves et al., 2007; $\alpha = .88$).

High opportunity-to-observe managerial behaviors. I averaged the experts' ratings from the pilot study to classify each rating dimension as *high* or *low opportunity-to-observe managerial behaviors*. Using a mean split, I coded high *opportunity-to-observe managerial behaviors* as 1 (*yes*) and the low *opportunity-to-observe managerial behaviors* as 0 (*no*). Because there was an uneven number of dimensions after removing the three dimensions with unacceptable levels of agreement, I classified the seven highest

dimensions as *high* and the remaining six as *low*. For peers, the dimensions of *Building/Mending Relationships* (*M* = 3.7)*, Decisiveness* (*M* = 3.6)*, Participative Management* (*M* = 3.6)*, Change Management* (*M* = 3.3)*, Career Management* (*M* = 3.1)*, Doing Whatever it Takes* (*M* = 3.1)*,* and *Being a Quick Study* (*M* = 3.0) were classified as *high opportunity-to-observe managerial behaviors* and the dimensions of *Confronting Problem Employees* (*M* = 2.1)*, Compassion and Sensitivity* (*M* = 2.1)*, Self-Awareness* (*M* = 2.5)*, Leading Employees* (*M* = 2.6)*, Resourcefulness* (*M* = 2.6)*,* and *Balance Between Personal and Work Life* (*M* = 2.6) were classified as *low opportunity-to-observe managerial behaviors.* For subordinates, the dimensions of *Putting People at Ease* (*M* = 4.3)*, Participative Management* (*M* = 3.9)*, Decisiveness* (*M* = 3.8)*, Doing Whatever it Takes* (*M* = 3.8)*, Change Management* (*M* = 3.8)*, Leading Employees* (*M* = 3.6)*, and Differences Matter* (*M* = 3.5) were classified as *high opportunity-to-observe behaviors* and the dimensions of *Self-Awareness* (*M* = 2.6)*, Career Management* (*M* = 2.7)*, Being a Quick Study* (*M* = 3.0)*, Confronting Problem Employees* (*M* = 3.0)*, Compassion and Sensitivity* (*M* = 3.1)*, and Resourcefulness* (*M* = 3.2) were classified as *low opportunity-to-observe behaviors*.

Extraversion-related managerial behaviors. I classified Benchmarks® dimensions into those that related to the focal manager's extraversion, and those that did not based on prior research (Center for Creative Leadership, 2004). This prior research examined the correlations between the focal manager's personality and the magnitude of peer and subordinate ratings for each rating dimension. Specifically, for peers, the dimensions of *Participative Management*, *Change Management*, *Doing Whatever it Takes*, *Decisiveness*, *Leading Employees*, *Confronting Problem Employees*, and

Compassion and Sensitivity were positively related to the focal manager's extraversion and therefore, were coded as 1 (*yes*), and *Resourcefulness, Being a Quick Study, Building and Mending Relationships, Balance Between Personal Life and Work, Self-Awareness, Putting People at Ease, Differences Matter*, and *Career Management* were unrelated to extraversion and were coded as 0 (*no*). For subordinates, the dimensions *Participative Management* and *Being a Quick Study* (negatively related) were coded as 1 (*yes*), or as being related to the focal manager's extraversion. *Resourcefulness, Doing Whatever it Takes, Decisiveness, Leading Employees, Confronting Problem Employees, Change Management, Building and Mending Relationships, Compassion and Sensitivity, Straightforwardness and Composure, Balance Between Personal and Work Life, Self-Awareness, Putting People at Ease, Differences Matter* and *Career Management* were coded as 0 (*no*), or as being unrelated to a focal manager's extraversion.

Personality dimensions. Focal managers completed two personality profiles: the Myers-Briggs Type Indicator (MBTI) (Myers & McCaulley, 1985) and the Fundamental Interpersonal Relations Orientation-Behavior (FIRO-B) (Schutz, 1957). Also, the focal manager's self rating on the *Self-Awareness* dimension of Benchmarks® was used as an additional personality indicator. The MBTI assesses Jung's (1971) theory of four different processes. Each of the four processes is divided into extremes, or pairs of personal preferences: extraversion-introversion, sensing-intuition, thinking-feeling, and judgment-perception. The MBTI identifies which trait of the pair an individual prefers in his or her daily actions. Extraverted (E) individuals tend to engage with their external environment whereas introverts (I) focus their attention on their inner environment. Those who have a sensing (S) type of perception are interested in what is concrete and

practical. In contrast, intuitive (N) types are more comfortable contemplating less observable processes such as hypothetical and symbolic scenarios. The thinking (T) and feeling (F) dimension assesses how a person makes a decision. Those who favor thinking processes arrive at decisions in a logical and rational manner, whereas feeling types use their values and feelings to make decisions. The fourth process measures whether people use judgment (J) or perception (P) to assess their external environment. Individuals who endorse the judgment type tend to organize and plan; they also arrive at decisions easily. Those who prefer perception processes, however, prefer to continue to receive information and remain receptive to alternative solutions.

Form M of the MBTI was used. This version is comprised of 93 forced-choice or dichotomous items where individuals reported whether a particular item does or does not indicate their preference (Myers & McCaulley, 1985). Each item reflects one of the eight possible types. The MBTI Form M is typically scored such that individuals are classified as falling into one extreme of each of the four pairs of types. Unlike other forms of the instrument (e.g., Form G) that were based on summative scoring, Form M is based on Item Response Theory (IRT) scoring (Briggs, Myers, McCaulley, Quenk, & Hammer, 1998). This method of scoring is associated with higher reliability and greater differentiation among types (e.g., fewer individuals receive tied trait preferences with Form M). The resulting feedback is often expressed through 1 of 16 possible combinations of preferences, such as ESFP, which represents a combination of extraversion, sensing, feeling, and perceiving preferences. In addition, the report for MBTI form M provides information regarding the strength of each preference. For each

dimension, an individual's preference score can range from 1 to 30; larger numbers represent stronger preferences.

The MBTI instrument has been extensively researched. MBTI scores have acceptable internal consistency (Briggs et al., 1998) and also relate to other well-established measures of personality. For instance, a study by Furnham, Moutafi, and Crump (2003) correlated the MBTI with the NEO PI-R (Costa & McCrae, 1992), which measures the Big 5 personality traits: extraversion, openness to experience, agreeableness, and conscientiousness. After controlling for gender and age, the NEO PI-R measure for extraversion had moderate to high correlations with the Myers-Briggs types E (r =.71) and I (r = -.72). In addition, they found that openness to experience related significantly to S (r = -.66) and N (r = .64), as did agreeableness to T (r = -.41) and F (r = .28), and conscientiousness to J (r = .46) and P (r = -.46).

For the purposes of the present research, I converted the eight continuous scores into four bipolar scales (i.e., EI, SN, TF, JP) based on findings from prior research (McCrae & Costa, 1989). A binary measure for each type was formed. Types E, S, T, and J were coded as 0 and I, N, F, and P were coded as 1. A continuous measure of these bipolar scales was also used. Lower negative numbers represent stronger preferences on E, S, T, and J whereas higher positive numbers represent stronger I, N, F, and P tendencies; the range for each scale is -30 to +30. Scores near zero represent an individual who does not have a strong preference on that particular trait pair. Based on the findings of Furnham et al. (2003), I tested hypotheses regarding extraversion using EI scores, conscientiousness using JP scores, and agreeableness using TF scores.

The second measure of personality, the FIRO-B (Schutz, 1957), measures how an individual interacts with others and is comprised of three dimensions: inclusion, control, and affection. Each dimension is also subdivided into expressed behavior (e), or how an individual acts toward others, and wanted behavior (w), or how an individual would like to be treated by others. The result is a three by two matrix. For example, the inclusion dimension assesses the extent that an individual wants to be included and includes others in interactions. The control dimension measures the extent that an individual controls others and desires to be controlled by others. The affection dimension assesses the extent that an individual expresses intimacy and forms personal relationships with others as well as desires others to form personal relationships with him or her.

The instrument is comprised of 54 Guttman-type items that vary in intensity and are measured with six-point scales. There is evidence that the scores on the FIRO-B are internally consistent and demonstrate acceptable levels of test-retest reliability (Schutz, Hammer, & Schnell, 2000). For the present research, scores for the inclusion, control, and affection dimensions were used. The scores for each dimension ranged between 0 and 18, with higher numbers indicating a greater interpersonal need in that particular area.

The focal manager's self-awareness was measured by his or her self rating on the *Self-Awareness* dimension of Benchmarks®. This dimension is the average of four items that assess managerial self-confidence and knowledge of personal strengths and weaknesses. The focal manager's self- rating on these items could range from 1 (*not at all*) to 5 (*to a very great extent*). A sample item is presented in Appendix B.

Organizational level of focal manager. Focal managers reported information about the hierarchical level of their jobs. *Organizational level* was coded as 1 = first level, 2 = middle, 3 = upper middle, 4 = executive, or 5 = top management.

Organizational tenure of focal manager. Focal managers reported in years the amount of tenure that they had at their current organization. There were 25 cases with tenure values that appeared to be miskeyed (i.e., negative numbers, tenure exceeding age in years, tenure greater than 70). For each of these cases, I recoded tenure as missing.

Rater source. The type of rater group was coded as either 0 for subordinates or 1 for peers.

Demographic characteristics. Focal managers and raters reported information about their demographic characteristics. This information included their *age*, *education*, *gender*, and *race*. The *age* of the focal managers and raters was reported in number of years. I recoded age data as missing from three raters that appeared to be miskeyed (e.g., ages less than 16). To capture *education*, participants reported their highest educational degree that they completed (coded as 1= high school, 2 = associate's degree, 3 = bachelor's degree, 4 = master's degree, 5 = professional degree/doctorate). The *gender* of the focal managers and the raters was coded as 0 for females and 1 for males. The *race* of participants was also coded dichotomously. Though respondents had the option to report whether they belonged to a number of racial and ethnic groups, the sample was predominantly Caucasian (82%), and therefore, the combination of multiple racial minorities did not occur with enough frequency to test for meaningful differences. Moreover, the primary purpose of the present research was to examine diversity within the group, not the specific effects of different racial groups. Other researchers have

similarly collapsed race into a dichotomous variable (e.g., Kirchmeyer, 1995; Ostroff et al., 2004). For this reason, raters who are not Caucasian were coded as 0 and those who are Caucasian was coded as 1.

Rater group diversity. The diversity of both peer and subordinate rater groups was calculated individually for the characteristics of age, education, gender and race. These variables were calculated for groups with demographic information from four or more group members. To calculate the diversity of continuous and ordinal demographic variables, the coefficient of variation was used. The coefficient of variation is recommended by Allison (1978) and is frequently used to represent diversity in research (e.g., Jackson et al., 1991; Jackson & Joshi, 2004; Klein et al., 2001; O'Reilly et al., 1989). The formula for the coefficient of variation (V) is:

$$\mathrm{V} = \frac{SD_x}{\overline{X}}, \tag{22}$$

where SD_x is the standard deviation of the rater group on variable X, and $\overline{X}$ is the overall mean for the rating group for variable X. This method was used for age (in years), and education. The coefficient of variation was calculated for each individual rating group. A score of zero indicates perfect homogeneity or similarity within a rating group on a given characteristic whereas larger numbers are representative of more diverse groups.

There are some shortcomings to the coefficient of variation. The coefficient of variation is most appropriate for use with ratio scales with real zero-points (Allison, 1978), but less so for interval-level data. Also, the values of V are related to the mean. For instance, two groups can have the same standard deviation, but the group with the higher mean, will have a lower deviation score (Bedeian & Mossholder, 2000). Moreover, V is impacted by sample size. Larger groups typically have greater variation,

and thus, values of *V* decrease with larger groups (Bedeian & Mossholder, 2000). However, there seems to be no preferred alternative to *V* that is without similar drawbacks.

To calculate the diversity of categorical variables, the proportional diversity measure was calculated. This method of calculating composition has been used in prior research (e.g., Randel, 2002; Rentsch & Klimoski, 2001). This measure was calculated by taking the proportion of women in the rater group to measure gender diversity. Race was treated similarly. Because I did not make directional hypotheses about composition (e.g., that higher agreement would be associated with a particular gender or race), I rescaled the proportion such that the greatest level of heterogeneity occurs when $P = .50$. Specifically, I converted the values of the proportional diversity measure such that .50 was the maximum value, similar to Klein at al. (2001). For instance, a value of .60 was recoded .40 and .90 was recoded .10. Therefore, a value of 0 represents complete homogeneity and .50 represents maximum diversity. This method of assessing composition has some advantages over other measures such as Teachman's (1980) index and Blau's (1977) which both yield scales with unequal intervals that can make theoretical interpretations difficult (e.g., Williams, 2004).

Rater acquaintance with focal manager. Participants reported how well they knew the focal manager on a four-point scale ranging from 1 (I hardly know this person) to 4 (I know this person extremely well). This item was averaged within rating groups to assess the degree that the rater group was acquainted with the focal manager. Higher values represent greater average levels of acquaintance. Although acquaintance was measured with just one item, it was aggregated at the group level. Other researchers

concluded that one-item measures can be adequately reliable when aggregated at the group level (Harter, Schmidt, & Hayes, 2002).

Within-source agreement. Earlier I presented the methods for assessing interrater agreement and I reviewed some of the strengths and weaknesses of each measure. Although these measures tend to correlate highly with one another (Roberson et al., 2007), I concluded that a_{wg} was the best method to assess agreement for the present study based on some of the advantages over r_{wg}. Specifically, r_{wg} indices can provide biased estimates of agreement for groups less than 10 and when the incorrect null distribution is used (e.g., Brown & Hauenstein, 2005). Perhaps the largest advantage of a_{wg} over r_{wg} is that a_{wg} takes into consideration how the potential variance of ratings can differ along different points of the rating scale, with the largest potential variance being at the scale midpoint. Thus, one consequence of using r_{wg} is that the mean rating tends to highly correlate with estimates of agreement (e.g., Brown & Hauenstein, 2005), whereas the location of the group mean has less impact on the values of a_{wg}. Given these advantages, I used a_{wg} to assess within-source agreement of MSF groups.

To demonstrate how a_{wg} was calculated for the present study, recall equation 15:

$$a_{WG} = 1 - \frac{2 * S_X^2}{\left[(H + L) * M - (M^2) - (H * L)\right] * \left[k/(k-1)\right]} \qquad (15)$$

For this study, k is the number of raters in the rating group; M is the mean of the particular rating dimension across a group of raters. In addition, H is equal to five, or the maximum possible value of the Benchmarks® rating scale; L is equal to one, or the lowest possible value of the Benchmarks® rating scale. Finally, S^2_x is the observed variance of the ratings for each rating group. I calculated values of a_{wg} separately for subordinate and peer groups on each of the 16 Benchmarks® rating scales. Thus, up to 32

indices of agreement were calculated for each focal manager. In addition, I averaged the agreement across the 16 scales to arrive at an average agreement for each rating group. Values of a_{wg} could range between -1.0 and 1.0, with higher values indicative of higher agreement.

Because a_{wg} is a relatively new measure, it has not been applied to MSF ratings. As the first study to compute a_{wg} for MSF rating groups, it was important to provide a comparison with other methods of assessing agreement that have been used in prior research (e.g., LeBreton et al., 2003). To do so, I calculated r_{wg} using LeBreton et al.'s (2003) methodology. I computed r_{wg} using two rating scenarios: a uniform distribution ($\sigma_E^2 = 2.00$) and a moderately skewed distribution ($\sigma_E^2 = 0.90$) using equation 10.

Sample selection criteria

Size of rater group. As mentioned previously, one potential drawback of calculating a_{wg} for MSF rating groups is that an adequate number of raters is needed to calculate all possible values of a_{wg}. Brown and Hauenstein (2005) recommended that groups should be comprised of at least four raters for a scale with five response options, which the Benchmarks® rating scale contains. Because the present study was the first to use a_{wg} to calculate agreement for MSF data, I followed these recommendations. The number of raters per rating group was assessed for each rating dimension. On average, 53% of peer and 56% of subordinate rater groups were removed per rating dimension due to inadequate group size.

Furthermore, I was not able to calculate agreement for groups with mean ratings that fell outside of the range of interpretable values which was determined by equations 16 and 17. Specifically, for groups of four with ratings on a five-point scale, I was only

able to calculate a_{wg} for group means that fell between 2.0 and 4.0. In addition, I determined the interpretable ranges for other group sizes (refer to Appendix C). Groups with mean values that fell outside of these ranges were treated as missing values. One exception was for groups with a mean equal to 1 or 5 (this occurred for an average of .4 % of peer groups and .8% of subordinate groups). For these groups, agreement was set to 1.0, or perfect agreement. In total, the remaining sample was just over one-third of the original sample (36% of peers and 34% of subordinates). On average, there were 11,723 peer groups and 10,159 subordinate groups per rating dimension with which I could test my hypotheses. Table 2 and Table 3 (for peers and subordinates, respectively) present the number and proportion of rater groups with four or more raters and with observed means for which a_{wg} could be calculated.

Missing data. It was important that the data set included the relevant variables to be able to properly test the hypotheses. For this reason, rater groups were excluded if they were missing Benchmarks® scale scores or if the number of ratings fell below the minimum of four. For demographic variables, I only calculated demographic composition indices if there were four or more values for the variables of age, gender, race, and education; groups with less than four values for a demographic trait were not included in the analyses. Because demographic information was missing for a number of groups, composition indices were only calculated for an average of 13,132 (39%) peer groups and an average of 13,097 (39%) subordinate groups. Personality variables were only available for a subset of 7,257 focal managers. Of the subset of data with focal manager personality data, I was able to calculate a_{wg} for an average of 2,469 (34%) peer groups and 2,076 (29%) subordinate groups.

Analyses. To test my hypotheses, I examined the correlations between within-source agreement and mean acquaintance (Hypothesis 1) and overall managerial effectiveness (Hypothesis 6). Multiple regression analyses were used to test hypotheses with multiple predicators (i.e., Hypothesis 2, Hypothesis 5a, and Research Question 1). Within each block of predictors, I entered variables simultaneously. For Hypothesis 5c, which predicted an interaction effect between personality and acquaintance, the personality predictor variables were entered in the first step to assess their relationship with within-source agreement. In the second step, mean group acquaintance was entered, and the cross-products of the personality and acquaintance variables were entered in the third step. I used *t*-tests to contrast levels of agreement among different rating sources (Hypothesis 3a) and rating dimensions (Hypothesis 4).

When testing all hypotheses, the criterion for statistical significance was $\alpha = .05$; I used two-tailed tests. Because hypotheses were tested using a very large sample, it was likely that many findings would be statistically significant but not necessarily practically significant. As such, effect sizes were classified according to Cohen's (1988) conventions. I interpreted a correlation of .10 as a small effect, a correlation of .30 as a medium effect, and a correlation of .50 as a large effect size. To determine the effect size from *t* tests, *r* was calculated using the following equation (Rosenthal, 1991):

$$r = \sqrt{\frac{t^2}{t^2 + df}} \tag{23}$$

where *t* is the *t* value from the *t*-test and df is the degrees of freedom from the *t*-test.

CHAPTER 4: Results

The correlation coefficients, means, and standard deviations for all predictor variables and a_{wg} indices for the 16 Benchmarks® dimensions are shown in Tables 4 and 5, for peers and subordinates, respectively. Examining the Pearson correlation coefficients revealed that correlations between the agreement indices for the 16 Benchmarks® scales and the predictor variables were very small to small in magnitude, although many were statistically significant due to the large sample size. The average intercorrelations among Benchmarks® a_{wg} indices were large (r = .51 for peers and r = .56 for subordinates), suggesting that rating groups who agreed about a focal manager on one rating scale, tended to agree on other scales as well.

Exploratory analyses. I expected that some of my independent variables would be related. For instance, the organizational level of the rater might relate to educational level as well as other demographic variables. Although I did find some indication of this, the relationships were not so strong that multicollinearity was a concern. Educational diversity and age diversity were the most highly correlated of the group demographic variables; $r\ (12,625) = .11$, $p < .001$ for peers and $r\ (12,599) = .12$, $p < .001$ for subordinates, suggesting that groups who were more homogeneous in age were also more homogeneous in terms of education. Also, Hypothesis 5c predicted that acquaintance would moderate the relationship between personality and agreement, but it is possible that personality variables related to acquaintance (e.g., extraverts may have greater acquaintance associated with them). I investigated this question and found that the relationships between the personality variables and acquaintance were small. The strongest relationship was found for the dimension of extraversion/introversion and

average group acquaintance (r (6,303) = -.07, $p < .001$ for peers and r *(5,876)* = -.10, $p < .001$ for subordinates, suggesting that there is a weak tendency for managers who are more extraverted to have higher levels of acquaintance with their raters.

In addition, although I did not hypothesize that the focal manager's demographic characteristics would impact within-source agreement, I regressed the 16 dimensions of within-source agreement onto the focal manager's gender, age, tenure, race, organizational level, and level of experience (see Table 6 for peers and Table 7 for subordinates). The purpose of these analyses was to determine whether the focal manager's characteristics should be used as control variables for subsequent analyses. Generally, the relationships between the focal manager's characteristics and within-source agreement were negligible. The focal manager's race, level of experience, and organizational level were generally unrelated to within-source agreement among peers and subordinates. Male focal managers had virtually identical agreement associated with them (M = .84) as female managers (M = .83); the correlations between gender and agreement ranged between -.01 and .05 (M = .04). I also found a very small trend suggesting that younger focal managers, on average, had somewhat higher within-source agreement than older focal managers with correlations ranging between -.06 and .02 (M = -.03). In addition, managers with longer organizational tenure rather than shorter organizational tenure had, on average, slightly higher agreement among their peers and subordinates; correlations ranged between -.02 and .07 (M = .02). Although I found some statistically significant relationships for the focal manager's gender, age, and tenure, the relationships were very small. Moreover, these relationships were not always directionally consistent (e.g., correlation coefficients were both positive and negative).

Regression models with all of the focal manager's characteristics did not account for more than 1% of the variance of within-source agreement, and for this reason, the focal manager's characteristics were not used as control variables when testing my hypotheses.

I performed exploratory analyses to assess the means, standard deviations, and normality of the predictor variables (see Table 8). In all cases, the kurtosis and skew indices were in the acceptable +/-2.0 range (Field, 2005). However, the most problematic set of predictors with respect to kurtosis and skew, were the binary MBTI scores which had an average kurtosis of -1.45. Fortunately, the continuous MBTI scores did not have this problem, with an average kurtosis of -.58. For this reason, I decided to test the personality hypotheses using the continuous MBTI scores rather than the binary MBTI types. In addition, histograms of the gender and racial diversity variables showed that rating groups clustered around a few discrete values. These proportional variables were a function of group size, and therefore, the possible values of these proportions were limited to only a few discrete values (e.g., .00, .25, .50 for groups of four). Because of this type of distribution, I decided to compare extreme groups using a t-test, in addition to performing regression analyses.

Comparison of a_{wg} and r_{wg}. This is the first study to calculate both a_{wg} and r_{wg} in a multisource feedback setting. As such, I compared the two indices in terms of their magnitudes, standard deviations, skew, and kurtosis which are presented for each of the Benchmarks® dimensions in Table 9 and 10 for peers and subordinates, respectively. I also compared the consistency of the indices and the extent that the magnitude of the indices related to the observed group mean and the size of the rater group. For all of these comparisons, the groups with a_{wg} are those that had at least four raters per rating

dimension and whose observed mean was inside the interpretable range, whereas r_{wg}^{-un} and r_{wg}^{-ms} were calculated for all rater groups (similar to LeBreton et al.'s (2003) methodology).

When comparing agreement magnitude, I expected that the mean values of r_{wg}^{-un} would exceed those of a_{wg}. This hypothesis was based on Brown and Hauenstein's (2005) demonstration that values of a_{wg} decreased as the group mean diverged from the scale midpoint, whereas values of r_{wg} remained constant irrespective of the observed mean. The average agreement among peers across all 16 Benchmarks® dimensions was .85 (*range* = .77 - .88) for a_{wg}, .85 (*range* = .80 - .89) for r_{wg}^{-un} and .69 (*range* = .61 - .76) for r_{wg}^{-ms}. For subordinates, the mean agreement across all 16 Benchmarks® dimensions was .81(*range* = .73 - .86) for a_{wg}, .83 (*range* = .78 - .88) for r_{wg}^{-un} and .65 (*range* = .57 - .75) for r_{wg}^{-ms}. Thus, for both rating sources, agreement levels were very similar for a_{wg} and r_{wg}^{-un}; but lower for r_{wg}^{-ms}. The lower level of agreement for r_{wg}^{-ms} is because this index accounts for agreement due to chance or response bias, whereas the other two indices do not account for such biases.

I examined the proportion of rater groups who had high levels of agreement and the proportion with low levels of agreement for each rating dimension (see Table 11). Brown and Hauenstein (2005) stated that rater groups with agreement equal or greater to .80 should be classified as having high agreement. I found that across all rating dimensions, an average of 74% of peers and 63% of subordinates had high levels of agreement based on this standard. In particular, there was the highest level of agreement for the *Resourcefulness* dimension (84% of peers and 77% of subordinates had high agreement). Low levels of agreement among rater groups, defined as groups with

agreement less than .60 (Brown & Hauenstein, 2005), were found for an average of 5% of peers and 10% of subordinates. The *Putting People at Ease* dimension had the lowest level of agreement for both peers and subordinates (16% of peers and 22% of subordinates had low levels of agreement). These fairly frequent occurrences of low agreement are particularly notable because it may not be appropriate to aggregate responses when agreement is less than .60 (Brown & Hauenstein, 2005).

Comparing the relative variation of the agreement indices was important because the purpose of this research was to predict variation in agreement among rating groups. Thus, higher standard deviations in the agreement indices would allow for more variation to be explained. For peer raters, the average standard deviation of agreement was .13 (*range* = .10 - .19) for a_{wg}, .16 (*range* = .12 - .20) for r_{wg}^{-un} and .27 (*range* = .22 - .31) for r_{wg}^{-ms}. For subordinates, the average standard deviation of agreement was .16 (*range* = .12 - .21) for a_{wg}, .18 (*range* = .13 - .22) for r_{wg}^{-un} and .29 (*range* = .24 - .33) for r_{wg}^{-ms}. On average, the variation among all agreement indices was low. The a_{wg} indices had the smallest average standard deviations, followed by r_{wg}^{-un}, and then r_{wg}^{-ms}. One implication of the low standard deviations of the agreement indices is that there is little variance that can be explained by the proposed predictors. This problem is especially a concern for the a_{wg} index, hence making it more difficult to find support for my research hypotheses using this index of agreement.

I also examined the distribution of the agreement indices by calculating skewness and kurtosis. The average skewness of agreement indices among peers was -1.79 for a_{wg}, -2.29 for r_{wg}^{-un} and -1.13 for r_{wg}^{-ms}. For subordinates, the average skewness of agreement indices was -1.68 for a_{wg}, -2.10 for r_{wg}^{-un} and -.95 for r_{wg}^{-ms}. These negative values

indicate that scores were concentrated at higher levels of agreement, which has been reported by other researchers (e.g., Mount, 1984). The r_{wg}^{-un} index exceeded the +/- 2.0 range that is generally deemed acceptable for skewness (Field, 2005), whereas a_{wg} and r_{wg}^{-ms} were within this acceptable range. The average kurtosis for agreement indices among peers was 5.32 for a_{wg}, 7.25 for r_{wg}^{-un} and .55 for r_{wg}^{-ms}. For subordinates, the average kurtosis of agreement indices was 4.65 for a_{wg}, 5.67 for r_{wg}^{-un} and -.03 for r_{wg}^{-ms}. These numbers indicate that the distributions of a_{wg} and r_{wg}^{-un} were leptokurtic, or that the distributions tended to be pointy. The kurtosis indices for both agreement indices were also larger than what is generally acceptable (i.e., +/- 2.0 range; Field, 2005); however, the kurtosis for r_{wg}^{-ms} was in the acceptable range. Based on the skewness and kurtosis of the agreement indices, r_{wg}^{-ms} was the most normally distributed index, followed by a_{wg}, and then r_{wg}^{-un}. The r_{wg}^{-ms} index was the most normally distributed index as a consequence of rescaling agreement to reflect a moderately skewed response bias.

Despite some of the differences in the distributions of the three indices, prior research suggests that the agreement indices are highly correlated (Roberson et al., 2007). Specifically, Roberson et al., using a Monte Carlo simulation, found that the correlation between a_{wg} and r_{wg} was .96. I found similarly high levels of consistency between a_{wg} and r_{wg} in the present study. I calculated the Pearson correlations between the indices for each rating dimension. The average correlation between a_{wg} and r_{wg}^{-un} was .98 for peers and .97 for subordinates and between a_{wg} and r_{wg}^{-ms} the average correlation was .94 for peers and .92 for subordinates. These findings suggest that the agreement indices are highly consistent despite some variations in magnitude.

One of the largest benefits of calculating a_{wg} instead of r_{wg} concerns the relationship of the agreement index to the group's mean rating. Brown and Hauenstein (2005) demonstrated that r_{wg} had a stronger association with the location of the group mean than a_{wg} (i.e., r_{wg}: r(89) = .63, $p < .05$; a_{wg} r(87) = -.03, $p > .05$) for job relevance ratings made by 27 experts. In their analyses, they reflected mean ratings that were below the scale midpoint so that the deviations were unidirectional. I replicated this analysis. First, mean ratings below the scale mean (3.0) were reflected. Then I compared correlations between the mean rating for each Benchmarks® dimension and a_{wg} and r_{wg}^{*un}. The average correlation between a_{wg} and the scale mean was $r = .11$ for both peers and subordinates (all but one relationship was significant at the $p < .01$ level). The average relationship between the mean rating and r_{wg}^{*un} was $r = .25$ for peers and $r = .32$ for subordinates (see Table 12); all relationships were significant at the $p < .01$ level. The relationship between the observed mean and agreement was stronger for r_{wg} than a_{wg}; however, the magnitude difference was considerably smaller than Brown and Hauenstein's (2005) findings.

I also examined the extent to which values of a_{wg} and r_{wg}^{*un} varied as a function of group size. Because the interpretable range increases for a_{wg} as rating groups get larger, I would expect that values of a_{wg} would generally decrease with larger groups. Indeed, a_{wg} was negatively related to group size, with an average correlation of $r = -.16$ for both peers and subordinates. Group size had a negligible relationship with the values of r_{wg}^{*un} (average $r = -.03$ for peers and $r = -.04$ for subordinates). In terms of agreement magnitude, I examined the level of agreement for groups larger than 10. I selected groups of 10 because Brown and Hauenstein (2005) claim that this is the smallest group for

which r_{wg} should be calculated, and it also provides a relatively wide interpretable range for a_{wg} (i.e., between 1.4 and 4.6). For groups of at least 10 peers, the mean agreement was .81 for a_{wg} versus .84 for r_{wg}^{-un}. Agreement among groups of 10 or more subordinates was .75 for a_{wg} and .80 for r_{wg}^{-un}. This pattern of lower agreement for a_{wg} is indicative of the fact that the interpretable range for the a_{wg} index increases for larger groups, and consequently a_{wg} accounts for lower possible variance at these extremes. Thus, the values of the two indices are less similar when calculated for larger groups because a_{wg} accounts for lower possible variance at rating scale extremes whereas r_{wg} does not.

The comparison of a_{wg} and r_{wg} indices suggests that there are some trade-offs among these indices of agreement. Compared to values of r_{wg}^{-ms}, a_{wg} is more restricted in range and less normally distributed. Thus, the distribution of a_{wg} posed some problems for testing hypotheses which require adequate levels of variance across rater groups (Cohen et al., 2003). In addition, a_{wg} could not be calculated for a large number of rater groups whose means were outside of the interpretable range. In contrast, one large benefit of a_{wg} is that it is less related to the observed group mean than both r_{wg} indices, and consequently, does not violate the regression assumption of homoscedasticity as the r_{wg} indices do. In conclusion, on one hand, using a_{wg} could be problematic because of its distribution and range restriction. On the other hand, r_{wg} indices are more influenced by rating magnitude and violate the assumption of homoscedasticity. Based on this analysis of agreement indices, I decided that the violation of the homoscedasticity assumption was unacceptable to properly test the hypotheses. Therefore, a_{wg} seemed to be the best available agreement index with which to test my hypotheses.

To address some of the shortcomings of a_{wg}, I investigated whether performing a transformation would ameliorate the range restriction problem. I performed a natural log transformation after eliminating negative scores and reversing agreement scores (Field, 2005). This transformation improved the kurtosis and skewness of the agreement indices. However, the transformed variables were still restricted in range as Cohen et al. (2003) suggested is often the case. Based on my findings and theoretical reasons for a non-normal distribution, I chose to report analyses that used a_{wg} as a dependent variable rather than the transformed a_{wg}; however, the conclusions regarding my hypotheses were the same regardless of whether I tested my hypotheses using a_{wg} or the transformed a_{wg}.

Tests of hypotheses. To test Hypothesis 1, which predicted that a rater group's mean level of acquaintance with the focal manager would predict within-source agreement among peers and subordinates, I computed Pearson bivariate correlations. Generally, the correlation coefficients were near zero, with a range in magnitude between -.06 to 0.0 for peers and -.04 and 0.0 for subordinates. Table 4 for peers and Table 5 for subordinates report these relationships for each Benchmarks® dimension. Although some of the correlation coefficients were statistically significant (11 dimensions were significant for peers and 10 dimensions were significant for subordinates), the relationships were not practically significant and did not approach weak effect sizes. Moreover, the trend was toward a negative relationship between agreement and acquaintance, which is directionally inconsistent with my hypothesis.

It is possible that the relatively low variation of the group acquaintance variable may have obscured the predicted relationships. I conducted supplemental analyses to contrast within-source agreement for groups with high and low group acquaintance. I

used an independent samples t-test to contrast rater groups who were more than one standard deviation below the group average for acquaintance with those whose level of acquaintance exceeded the group average by more than one standard deviation (see Table 13 for peers and Table 14 for subordinates). Among peers, I found that the low acquaintance groups were significantly different for eight Benchmarks® dimensions. For seven of these eight dimensions, groups with low acquaintance had higher levels of agreement than the high acquaintance groups. Similarly, I found that among subordinates, there were significantly higher levels of agreement associated with the low acquaintance groups, rather than the high acquaintance groups for 6 of the 16 dimensions. The effect sizes for these relationships were very small (*r*s ranged between .04 to .08 for the significant relationships). However, I found consistent and significant patterns for the dimensions of *Confronting Problem Employees*, *Balance Between Personal and Work Life, and Self-Awareness* for both peers and subordinates. These three dimensions were also rated as being *low opportunity-to-observe* dimensions for both peers and subordinates (with the exception of *Balance Between Personal and Work Life* for peers, which was not classified due to low interrater agreement).

One possible reason for higher agreement being associated with lower levels of acquaintance for *low opportunity-to-observe* dimensions is that raters used stereotypes to make their ratings in the absence of adequate acquaintance and first-hand knowledge of the behavioral domain being rated. To examine this possibility, I performed a series of post-hoc ANOVAs to examine the relationship between rating groups that had high versus low acquaintance with the focal manager and the focal manager's gender in predicting agreement on the *Confronting Problem Employees*, *Balance Between Personal*

and Work Life, and *Self-Awareness* dimensions (see Table 15 for peers and Table 16 for subordinates). There are theoretical reasons why these three dimensions could be associated with gender stereotypes. Men, who are typically thought to embody more agentic traits may be perceived to be better at confronting employees, whereas women who are often associated with communal traits may be perceived to be better at balancing their work and family life (Heilman, 1995; Heilman, Block, Martell, & Simon, 1989). In addition, there is evidence that women tend to be more self-aware than men in assessing their managerial competence (Fletcher, 1999). As expected, I found significant main effects for acquaintance level predicting within-source agreement. I also found a positive main effect for focal manager gender predicting peer agreement for the *Balance Between Personal and Work Life,* $F(1, 2796) = 26.67$, $p < .001$, $\eta^2 = .10$ and *Self-Awareness,* $F(1, 3504) = 6.66$, $p < .01$, $\eta^2 = .04$ dimensions. In both cases, men had higher levels of agreement associated with them when compared to their female counterparts; this finding is consistent with exploratory analyses that indicated that raters had slightly higher agreement when rating a male manager rather than a female manager. However, I did not find any significant interactions between level of acquaintance and focal manager gender.

Hypothesis 2 predicted that the diversity of the rater group in terms of gender, race, education, and age would be negatively related within-source source agreement. To test this hypothesis, I entered these four predictors into the regression model separately for peers and subordinates for each dependent variable. The findings are shown in Tables 16 and 17 for peers and subordinates, respectively. For peers, racial diversity and gender diversity were small but significant predictors of agreement for each of the 16 rating dimensions and age diversity was significant in predicting 12 indices. Diversity in

education only predicted one dimension of agreement. For subordinates, racial diversity and age diversity significantly predicted all 16 rating dimensions, and educational diversity was significant in predicting 13 indices. Diversity in gender among subordinates only significantly predicted agreement for three dimensions. For all significant findings, I found that higher levels of diversity were associated with lower levels of agreement, as hypothesized. However, the effect sizes of these relationships were very small. At most, the R^2 of the regression model only accounted for 1% of the variance in within-source agreement.

The distributions of gender and racial diversity variables were not normal; values clustered around a few discrete proportions. Therefore, regression analyses using these variables may have resulted in inaccurate estimates of the relationship between agreement and the group composition of these variables. Instead, I felt that it was more appropriate to contrast groups that had high versus low diversity in terms of gender and race. Specifically, I contrasted groups that were more than one standard deviation below the group average for gender and racial composition with those whose level of gender and racial composition exceeded the group average by more than one standard deviation using an independent samples t-test. I found that groups with low diversity in terms of gender had significantly higher agreement for 15 rating dimension among peers (see Table 19) and for 8 dimensions among subordinates (see Table 20). Also, I found that peer and subordinate groups who were less racially diverse had higher agreement than more diverse groups for 15 of the 16 rating dimensions (see Tables 21 and 22, respectively). Thus, these supplemental analyses further support a directionally consistent relationship between agreement and group composition in terms of gender and race,

however, the size of these effects were still very small, with *r*s ranging between .03 and .08 for significant t-tests.

Hypothesis 3a stated that agreement would be higher for peers rather than subordinates. To test this hypothesis, I performed a dependent *t*-test between the agreement of peers and subordinates for each Benchmarks® rating dimension. Tenure information was not available for raters, and therefore, I was unable to control for rater tenure before performing this analysis as I had originally planned. Despite this limitation, I found support for this hypothesis. On average, peers had small but significantly higher levels of within-source agreement (M = .85, SD = .03) than subordinates (M = .81, SD = .03), ($t(15) = -13.87$, $p < .001$; the effect size of this relationship was large (r = .96).

Hypothesis 3b stated that rater group source would moderate the relationship between demographic characteristics and agreement such that the relationship between demographic diversity and agreement would be stronger for subordinates than for peers. However, because I did not find practically significant relationships between demographic characteristics and agreement for either rater source, I did not test for a moderating relationship between demographic characteristics and agreement.

Hypothesis 4 stated that agreement would be higher for *high opportunity-to-observe* rating dimensions than for *low opportunity-to-observe* rating dimensions. To test this hypothesis, I contrasted the mean within-source agreement for the *high opportunity-to-observe* dimensions with the *low opportunity-to-observe* rating dimensions using an independent *t*-test. For peers, within-source agreement for *high opportunity-to-observe* dimensions (M = *.86*, SD = .02) was not significantly higher than agreement for *low opportunity-to-observe* dimensions (M = *.84*, SD = .04), $t(11) = 1.04, p > .05)$; although

it did represent a medium effect size (r = .30). For subordinates, within-source agreement for *high opportunity-to-observe* dimensions (M = *.83*, SD = .01) was not significantly different from the agreement associated with *low opportunity-to-observe* dimensions (M = *.81*, SD = .03), $t(11) = .83, p > .05, r = .24$), however, the relationship did approach a medium effect size. Although the t-tests were not statistically significant for these analyses, the power of this statistical analysis was very low because so few dimensions were contrasted. Thus, it is notable to mention that the means were directionally consistent with my hypotheses and the effect sizes were medium in magnitude.

Hypothesis 5a predicted that the focal manager's level of extraversion, agreeableness, and conscientiousness would positively relate to within-source agreement. To test this hypothesis, I entered the focal manager's MBTI scores for the extraversion-introversion, thinking-feeling (Agreeableness), and judging-perceiving (conscientiousness) dimensions simultaneously in a regression equation to predict peer and subordinate agreement (see Table 23 for peers and Table 24 for subordinates). Although the personality traits of the focal manager significantly predicted agreement among peers for five Benchmarks® dimensions and two dimensions for subordinates, the R^2 for these regression models ranged between .001 and .006, suggesting that the relationship between the focal manager's personality and within-source agreement has little practical significance. One trend, despite these small effects was that focal managers who endorsed the Thinking (low Agreeableness) style had higher levels of agreement among peers for the *Doing Whatever it Takes*, *Leading Employees*, and *Confronting Problem Employees* dimensions. The effect sizes for these three dimensions were extremely small and counter to my hypothesis that Agreeableness would positively relate

to agreement. The personality traits of EI (extraversion) and JP (conscientiousness) did not show any consistent relationships with agreement across the Benchmarks® dimensions. Moreover, I conducted supplemental analyses to examine if there were differences in the levels of within-source agreement for managers who were on the extreme ends of these personality dimensions, but I did not find any notable patterns.

Hypothesis 5b stated that the relationship between extraversion and agreement would be stronger for rating dimensions that have previously shown to relate to the focal manager's level of extraversion. I did not test this hypothesis because I found only one statistically significant relationship between extraversion and agreement out of 32 separate analyses.

Research Question 1 asked whether other personality traits related to within-source agreement. To test this hypothesis, I entered the focal manager's MBTI Sensing-Intuition scale and the FIRO-B inclusion, control and affection scales simultaneously in a regression equation to predict peer and subordinate agreement (see Table 25 for peers and Table 26 for subordinates). Though the personality variables did significantly predict agreement for 8 of 16 dimensions for peers and 3 of 16 dimensions for subordinates, the relationships were very small and inconsistent across rating dimensions. At most, the R^2 for these regression models accounted for only 1% of the variance in within-source agreement. The focal manager's need for inclusion was positively related to agreement for six of the dimensions for peers and two dimensions for subordinates. The focal manager's need for affection was negatively related to agreement for five rating dimensions among peers, but was unrelated to agreement among subordinates. The focal manager's need for control was not related to agreement among peers, but was negatively

related to subordinate agreement for three dimensions. Focal managers who were sensing types had higher levels of agreement associated with them among their peers for four rating dimensions, but this trait was unrelated to agreement among subordinates. I also examined the relationship between the target manager's self-awareness and peer and subordinate agreement. The correlation coefficients between the target manager's self ratings on the *Self-Awareness* dimension and agreement were near zero, with a range in magnitude between -.01 to .02 (M = .00) for peers and -.02 and .01 (M = -.01) for subordinates

Hypothesis 5c predicted that the relationship between personality and agreement would be moderated by the extent that raters were acquainted with the focal manager, such that personality would have a stronger impact in predicting agreement with lower levels of acquaintance. To test this hypothesis, I added the focal managers' MBTI scores in Step 1. I entered the rater group acquaintance with the focal manager in Step 2. In Step 3, I entered the cross-products of personality scores and mean acquaintance. Tables 26 and 27 (for peers and subordinates, respectively) summarize the findings of regressions predicting the averaged a_{wg} across all 16 Benchmarks® dimensions. These models did not indicate any significant interaction effects. Similarly, analyses were repeated for each individual rating dimension for peers and subordinates without finding evidence of an interaction between personality and acquaintance and thus, these 32 analyses are not reported in the interest of conserving space.

I found support for Hypothesis 6, which predicted a positive relationship between the focal manager's overall managerial effectiveness rating and within-source agreement for peers and subordinates. The correlations between managerial effectiveness and

agreement (a_{wg}) ranged between .00 and .11 (M = .06) for peers and .00 and .08 (M = .06) for subordinates (see Table 4 for peers and Table 5 for subordinates). With the exception of the *Balance Between Life and Work* dimension, all other dimensions were significantly correlated with overall effectiveness at the $p < .01$ level. In particular, agreement among peers on the *Self-Awareness* (r = .11) and the *Doing Whatever it Takes* (r = .10) dimensions were the most highly correlated the supervisor effectiveness ratings. Among subordinates, agreement on the *Resourcefulness* (r = .08) and the *Doing Whatever it Takes* (r = .08) dimensions were the most highly correlated with the supervisor's overall effectiveness ratings.

Research Question 2 asked about the relative strength of the predictors tested. To answer this question, I planned to perform relative weights analyses. However, because I did not find strong support for my individual research hypotheses, comparing the strength of these predictors was not necessary.

In addition to my proposed analyses, I also conducted some supplemental analyses to assess the extent that some target managers were more judgable than others. I found that the intercorrelations among the agreement indices for each Benchmarks® dimension were relatively high (α = .94 for peers and α = .95 for subordinates). This finding indicates that raters consistently agreed on target managers from one behavioral domain to another. I was also curious whether peers and subordinates were consistent in the extent that they agreed when rating a particular target. I examined the intercorrelations between the a_{wg} indices for peers and subordinates. On average, the correlation between peer and subordinate agreement on each corresponding Benchmarks® dimension was r =.08, with a range between .05 - .12. This finding does

suggest that there is a small relationship between peer and subordinate agreement, or that peers and subordinates tended to agree on the same focal managers. In particular, the correlation between peer and subordinate agreement exceeded .10 for the dimensions of *Resourcefulness* (r = .12), *Doing Whatever it Takes* (r = .10), *Building and Mending Relationships* (r = .10), and *Differences Matter* (r = .11).

I carried out a number of additional analyses that are not reported, and which did not affect the overall conclusions regarding the predicted relationships. Specifically, I examined the relationships between the predictors and three additional sets of agreement indices: (1) the natural log transformation of a_{wg}, (2) r_{wg}^{-un}, and (3) r_{wg}^{-ms}. I tested hypotheses with a transformed a_{wg}, because of problems with range restriction and skew. However, this transformation did not solve the issue of range restriction (Cohen, 2003), and would hinder the interpretation of agreement among rater groups. I was also concerned that removing a large portion of rater groups from my sample when calculating a_{wg} due to uninterpretable ranges could potentially affect the generalizability of my findings. For this reason, I performed all analyses with r_{wg}^{-un} and r_{wg}^{-ms} as my dependent variables. These two dependent variables had similar relationships with the predictor variables as a_{wg}, and thus reporting the results of these analyses would be redundant.

CHAPTER 5: Discussion

The aim of the present research was to examine when rater groups are likely to agree on their multisource feedback ratings of a focal manager. Kenny's (1991) WAM of consensus was used as a framework to identify predictors of agreement, which included characteristics of the rater group, the ratee, and the dimension being rated. I found some support for my research hypotheses, including the finding that more effective managers tended to have higher levels of agreement associated with them than less effective managers. I also found that peers tended to agree with one another more than did subordinates.

Although I found patterns consistent with my hypothesis that diversity in demographic characteristics would negatively relate to within-source agreement, the effects were very small. There were hypotheses for which I failed to find support. I did not find a relationship between the opportunity to observe managerial behaviors and within-source agreement. In addition, rater group acquaintance with the focal manager and the focal manager's personality traits did not relate to within-source agreement. Table 29 provides an overview of findings for each hypothesis.

The present research provided evidence that rater groups generally have high within-source agreement, which has been disputed by other researchers (e.g., Greguras & Robie, 1998). In addition, I was the first to assess agreement among MSF rater groups using the relatively new a_{wg} index. I also used the empirically developed Benchmarks® multisource feedback instrument and analyzed MSF data from a very large sample of focal managers, and thus, the results of the present research should be highly generalizable to other MSF instruments.

Level of agreement. The present study was the first to assess within-source agreement using a_{wg} in a MSF context. On average, agreement was generally above Brown and Hauenstein's (2005) .80 criterion, or the level that represents a high level of agreement. Specifically, I found that about three-quarters of peer groups and almost two-thirds of subordinate groups exceeded the .80 criterion. This finding is somewhat contradictory from prior research studies that found low levels of rater group agreement using other statistical indices (e.g., Greguras & Robie, 1998).

However, it is worth noting that I found that weak and unacceptable levels of agreement ($a_{wg} < .60$) occurred for an average of 5% of peer and 10% of subordinate groups. Furthermore, some rating dimensions had more frequent incidences of low agreement than did others. In particular, for the *Putting People at Ease* dimension, 16% of peer groups and 22% of subordinate groups had unacceptable levels of agreement. Perhaps certain types of managerial behaviors, such as those captured in the *Putting People at Ease* dimension are more subjective or difficult to observe, and subsequently, raters agree with one another less often in their ratings. These relatively high incidences of unacceptable agreement are notable. Brown and Hauenstein (2005) stated that it is not appropriate to aggregate responses with values of a_{wg} less than .60. Thus, it is questionable whether an averaged feedback score is the most appropriate representation of rater feedback for these types of managerial behaviors.

Review of the research findings. One of the most notable findings of the present research was that there were higher levels of agreement associated with effective focal managers than ineffective managers. To the best of my knowledge only one prior study has examined the link between rater agreement and leadership ratings (e.g., Feinberg et

al., 2005). Although Feinberg et al.'s study found that raters had higher levels of agreement when rating more transformational leaders, the study's design was flawed because agreement and ratings of leader effectiveness were obtained from the same rating source. In addition, agreement was assessed using r_{wg}, which has been shown to be related to the group mean (Brown & Hauenstein, 2005). However, the present research did not have these methodological weaknesses; ratings of the focal manager's effectiveness were made by the focal manager's supervisor, and therefore were independent from peer and subordinate agreement. Moreover, the present research used a_{wg}, which is not as highly correlated with the location of the group mean as is r_{wg}.

Despite the differences from Feinberg et al.'s (2005) study, the results were similar. I found significant and directionally consistent relationships between the supervisor's rating of overall effectiveness and all indices of agreement among peers and subordinates, with the exception of the *Balance Between Work and Life* dimension; some of these relationships approached a small effect size. Peer raters had correlations that were at least .10 between overall effectiveness and the dimensions of *Doing Whatever it Takes* and *Self-Awareness.* These findings suggest that focal managers who are seen as generally effective by their supervisors tend to have peers who agree on their tenacity, vision, and ability to effectively seek and apply feedback from others. This explanation is consistent with Kenny's (1991) WAM which stated that higher agreement should occur when rating a more consistent person. Effective managers are likely to be more consistent because they have less variable track records than less effective managers who may perform well in some situations, but falter in others.

Another notable finding of the present research was the relative agreement of peers and subordinates. Other researchers have discussed ratings vis-à-vis the relative abilities of peers versus subordinates (e.g., Bates, 2002; London & Wohlers, 1991), concluding that subordinate raters may be less experienced or qualified to provide feedback ratings. I found some support for this notion; peers had higher levels of agreement than subordinate raters. This finding raises the question of whether subordinate ratings have more error than peer ratings. Perhaps agreement was lower among subordinates because they were less experienced than peers in providing high quality ratings or they may not have observed their supervisors performing relevant managerial behaviors as often as did peer raters. Also, some subordinates may be uncomfortable providing upward feedback regarding their supervisor's managerial abilities (London & Wohlers, 1991). These are all possible reasons why I found a greater tendency for peers to agree with one another than subordinates.

I generally found patterns consistent with my hypothesis that diversity in demographic characteristics would negatively relate to within-source agreement, although the effects were very small. Specifically, I found consistent relationships between within-source agreement and age, gender and racial diversity for peer raters and age, racial, and educational diversity for subordinate raters. In addition, when I contrasted extreme groups on gender, I found that agreement for eight rating dimensions was related to gender composition for subordinate groups, or that there was somewhat lower agreement among more diverse groups in terms of gender. Generally, there was no relationship between educational diversity and peer agreement. Incidentally, peer groups tended to be more homogeneous in educational degrees than subordinates, which may be

one reason for the lack of relationship between educational diversity and peer agreement. Again, the relationships I found were consistent with this hypothesis, but in practical terms, these relationships were very small.

I hypothesized that there would be higher levels of agreement when rating behaviors for which raters have a greater opportunity to observe. This hypothesis was not supported. However, the patterns of agreement and observability were consistent across peers and subordinates. Although agreement was slightly higher among the *high opportunity-to-observe* dimensions rather than the *low opportunity-to-observe* dimensions, this difference was not statistically significant. However, I had very low power to detect differences between the two groups of dimensions.

I expected that rater groups who were better acquainted with the focal manager would have higher levels of agreement than groups with lower levels of acquaintance, as has been found by some prior research (Rothstein, 1990). One reason for the lack of a relationship between agreement and acquaintance could be because a vast majority of the raters in the Benchmarks© sample had adequate levels of knowledge regarding the focal manager. This scenario is particularly plausible given that prior research has found that raters can make reliable and accurate ratings of a target with very low levels of acquaintance (Funder & Colvin, 1988; Kenny & Albright, 1994; Rothstein, 1990). Moreover, both peer and subordinate groups reported having a relatively high level of acquaintance with the focal manager (an average acquaintance level of 3.1 on a scale from 1 to 4).

Despite the general lack of a relationship between acquaintance and agreement, comparisons with high and low acquaintance groups revealed very small but consistent

negative relationships between acquaintance and agreement that were counter to my hypothesis. Specifically, I found evidence that groups with low levels of acquaintance tended to have higher levels of agreement than groups with high levels of acquaintance. Despite the small effect size, it is interesting to note that the three dimensions that had consistent relationships across peers and subordinates, *Confronting Problem Employees*, *Balance Between Personal and Work Life*, and *Self-Awareness*, were also classified as *low opportunity-to-observe* dimensions in the pretest for both groups (with the exception of *Balance Between Personal and Work Life* for subordinates, which was removed from the analysis due to low interrater agreement). Thus, these dimensions represent types of behaviors that are generally difficult to observe. Perhaps rater groups were more likely to apply stereotypes, thereby inflating agreement, when rating an unfamiliar focal manager on behavioral domains for which they had little opportunity to observe the focal manager performing. Another explanation for the relationship between lower levels of agreement among high acquaintance groups is that raters who are highly acquainted with target managers may have highly individualized relationships that are expressed through disparate ratings (Kenny & Albright, 1987; London & Smither, 1995). Thus, lack of agreement among high acquaintance groups could reflect the use of more specific or individualized behavioral cues rather than the more general surface-level cues that are used by raters with lower levels of acquaintance.

I did not find any indication that the focal manager's personality traits related to the level of agreement among raters. Although I found a slight trend that linked low levels of Agreeableness to higher levels of agreement, the relationships were very small and the pattern only occurred for 4 out of 32 Benchmarks® dimensions. Moreover, these

relationships were counter to my prediction. The exploratory personality traits that I examined also had very small effects, which were inconsistent across the dependent variables. Moreover, I did not find support for an interaction between the level of acquaintance and personality traits.

Methodological implications. My comparison of a_{wg} and r_{wg} suggests that there are some trade-offs to consider when choosing a method for assessing agreement. At first glance, a_{wg} was highly consistent with r_{wg}^{-un} and r_{wg}^{-ms}, which has been shown previously (Roberson et al., 2007). The magnitudes, ranges, and standard deviations were highly consistent across the a_{wg} and r_{wg}^{-un} indices. Though r_{wg}^{-ms} was consistent with a_{wg}, it had a smaller mean, a larger standard deviation and was more normally distributed than a_{wg} as a function of being rescaled to account for response biases. Despite these similarities, proponents of a_{wg} cited a number of benefits of a_{wg} over r_{wg} (e.g., Brown & Hauenstein, 2005), including a lack of scale dependency and that fewer raters are needed to calculate a_{wg} than r_{wg}. I evaluated these claims and found mixed results.

First, I compared the two indices in the extent that they were confounded with the observed mean. Unlike Brown and Hauenstein's (2005) comparison of a_{wg} and r_{wg} for job relevance ratings of situational judgment items, I did not find a strong relationship between r_{wg} and the observed mean (i.e., $r = .63$; Brown & Hauenstein, 2005). Instead, I found the relationship was of medium strength ($r = .25$ for peers and $r = .32$ for subordinates). Moreover, I found a weak relationship between a_{wg} and the observed mean ($r = .11$) whereas Brown and Hauenstein found no relationship. A possible reason for finding a weaker relationship between r_{wg} and the observed mean in the present research is that there was some range restriction in the agreement indices, which may have

downwardly biased the relationship between the mean and r_{wg}. However, range restriction does not explain why I found evidence of a relationship between a_{wg} and the mean rating, albeit a weak one.

Being able to calculate a_{wg} for groups of 4 raters rather 10 (as is recommended for r_{wg}) was not without its problems. Brown and Hauenstein (2005) did not fully explain the impact of small group size on a_{wg} calculations. The size of the rater group determines the range for which a_{wg} can be calculated. In the current study, this meant eliminating groups of four with observed means that were less than 2.0 or more than 4.0, groups of five with observed means that were less than 1.8 or more than 4.2, and so forth. Among rater groups that had four or more raters, a considerable portion (32% of peers and 40% of subordinates) were removed because their group means fell into an uninterpretable range.

The less obvious implication of calculating a_{wg} for small groups is that the lack of scale dependence for a_{wg} is not fully realized for small groups. Although the formula for a_{wg} downwardly adjusts the possible variance as the observed mean moves further from the scale midpoint, this benefit is moot if there is a large portion of the rating scale for which a_{wg} cannot be calculated. Put another way, values of a_{wg} and r_{wg}^{-un} are identical when the group's observed mean is equal to the scale midpoint (Brown & Hauenstein, 2005). However, values of a_{wg} become smaller relative to r_{wg}^{-un} when the group's mean moves closer to either scale extreme because the formula for a_{wg} accounts for the decreased variability at the scale extremes. The scale extremes are where there should be the greatest divergence between a_{wg} and r_{wg}^{-un}, and are precisely the areas outside of the interpretable range. This problem has the largest impact on the calculations for smaller groups.

Thus, it is not surprising that I found the average magnitude of a_{wg} to be very similar to r_{wg}^{-un} (.85 versus .85, respectively for peers and .83 versus .81 respectively for subordinates). There were some rating scenarios for which a_{wg} and r_{wg}^{-un} were less similar. I found that values of a_{wg} decreased with larger groups, whereas values of r_{wg}^{-un} did not relate to group size. The relationship between group size and magnitude of agreement occurred because larger groups have larger interpretable ranges, and thus, the maximum possible variance decreases as the group's mean moves further from the scale midpoint. Therefore, the benefits of a_{wg} over r_{wg}^{-un} are only likely to be realized for groups that are relatively large. Perhaps Brown and Hauenstein's (2005) guidelines for calculating agreement for groups one less than the number of response options are not conservative enough. In particular, many group means are likely to fall into an uninterpretable range in situations similar to MSF ratings – where rating groups are small and observed means approach either the lower or upper rating boundary.

This critique is not intended to discourage the use of a_{wg}; however, one should consider the research question and the rating scenario when choosing an agreement index. When using agreement as a dependent variable, as in the present research, a_{wg} is advantageous over r_{wg} in that it does not violate the regression assumption of homoscedasticity. However, for other research purposes, such as verifying levels of agreement in order to justify aggregation (e.g., Klein, et al., 2001), the possibility of not being able to calculate a_{wg} for a large portion of rater groups may be unacceptable. In such cases r_{wg} may be the preferred and more appropriate index of agreement.

Implications for theory. I applied Kenny's (1991) weighted-average model of consensus as a framework for identifying conditions that relate to rater agreement. The

goal was to better understand how attributes of the focal manager, the rater group, and the content domain being rated related to within-source agreement. Although I failed to find robust predictors of agreement across all of these variables, I did find some consistent patterns that suggest potential contributions to theory.

I found evidence that some focal managers were more judgable, or that raters agreed about some focal managers more than others. I found significant correlations among the 16 agreement indices for both peers and subordinates, which suggest that rater groups who agree on one rating dimension also agree on others. Moreover, I found a small but significant correlation between peer and subordinate agreement on the same rating dimension. This finding indicates that regardless of rating source, raters tend to agree on the same focal managers. Moreover, similar to prior research (e.g., Feinberg et al., 2005), my findings indicate that a focal manager's effectiveness relates to his or her judgability. Specifically, I found that focal managers who were rated as more effective by their supervisors also had higher levels of peer and subordinate agreement. This relationship between effectiveness and rater agreement may indicate that it is easier to rate a very high performing focal manager who consistently delivers excellent results, compared to a mediocre manager who has a more variable track record. Thus, the variable track record of less effective managers may increase the level of rating complexity, which ultimately results in higher disagreement among raters. This explanation is consistent with Kenny's (1991) WAM which stated that agreement among raters should be higher when rating a target who is behaviorally consistent.

Alternately, there are other possible reasons for the relationship between overall managerial effectiveness and rater agreement. Raters may be aware that these highly

effective managers are labeled as 'high potentials' by their supervisors or organization. The elevated status of these high potential managers may have activated heuristics that lead the raters to inflate their ratings. However, this explanation is more likely to impact the magnitude of the ratings, rather than the agreement among raters. Moreover, a_{wg} has been shown not to correlate with the mean rating to the extent that other agreement indices do (Brown & Hauenstein, 2005), and thus, the use of a_{wg} to assess agreement should minimize this possible confound.

Another area of theoretical contribution of this study concerns how characteristics of the rater group may predict agreement. I hypothesized that the demographic diversity of the rater group would negatively relate to within-source agreement. The basis for this hypothesis was twofold. First, the findings of prior research suggested that individuals who share similar backgrounds and experiences in the workplace are apt to interpret the behaviors of others more similarly than more diverse rating groups (e.g., Deal, 2005; Rentsch & Klimoski, 2001). Second, groups who share similar characteristics may be more likely to communicate about the focal manager with one another as compared to more diverse groups (Smith et al., 1994; Zenger & Lawrence, 1989). Although the patterns of agreement that I found were directionally consistent with this prediction for racial diversity, gender diversity, and age diversity among peers and racial diversity, age diversity, and educational diversity among subordinates, the relationships were very weak. Based on these findings I would conclude that the demographic composition of the rater group does not appear to have a substantial impact on within-source agreement, at least in these data. Because prior research found that deeper levels of diversity were better predictors of group outcomes than surface-level demographic traits (e.g., Harrison

et al., 1998), it is possible that more direct tests of the underlying processes for which demographic variables are proxies may better predict within-source agreement.

Another characteristic of the rating group that I examined was the mean level of acquaintance. The small trend that I found between lower levels of acquaintance and higher levels of agreement suggests that the theorized positive relationship between rater acquaintance and interrater agreement may not be as simple as the results from prior research have suggested (e.g., Rothstein, 1990). In particular, it may be useful to consider how the opportunity to observe the content domain interacts with the rater's acquaintance with the target manager. In addition, a number of researchers have noted that readily observable behaviors and clearly defined performance standards result in higher quality ratings than vague or poorly defined standards (e.g., Guion, 1998; Landy & Farr, 1980; Latham & Wexley, 1977). Thus, if there are some rating dimensions that are inherently less concrete or more difficult to observe, it makes sense that raters who are not well acquainted with a focal manager may rely on some sort of stereotype. This explanation is similar to the lens model (Brunswik, 1952; Schmitt et al., 1986) which states that inflated reliability or agreement may be the result of raters applying information that is unrelated to performance, such as shared stereotypes or biases. Further examination of the relationship between the content domain being rated and level of acquaintance may help to identify situations where interrater agreement does not logically reflect accuracy, but instead is an indicator of bias.

Implications for practice. I found some evidence that raters within peer and subordinate groups tend to agree with one another in their ratings of focal managers. This finding confirms one of the key assumptions underlying the use of multisource feedback -

that agreement among raters from the same source is adequately high (Borman, 1997). However, I also found that there was still a sizeable proportion of rater groups, approximately 5% of peers and 10% of subordinates, who failed to agree with one another at a level needed to aggregate responses (Brown & Hauenstein, 2005). This lack of agreement was considerably more frequent for the rating dimensions *Putting People at Ease, Self-Awareness, Balance Between Personal and Work Life.* These three rating dimensions represent personal traits and behaviors that may be more difficult for raters to observe than more concrete dimensions such as *Resourcefulness*. One implication is that information about rater agreement could be used in the development of multisource feedback instruments. For example, practitioners should consider minimizing their use of rating dimensions for which there is likely to be a high level of disagreement within the rater group.

Another related implication of this research is that when there is low agreement for a rating, the mean rating may not accurately convey the feedback that the separate raters provided (Chan, 1998). For example, if two raters rated the focal manager very low on *Leading Others* whereas two other raters made very high ratings, the mean of the four ratings will obscure the underlying bimodal distribution (Chan, 1998). Therefore, the potential effectiveness of MSF feedback is questionable for groups with low levels of agreement. Moreover, because I found some indication that weaker focal managers tended to have lower levels of agreement associated with them, it is possible that the effectiveness of MSF may be systematically lower for less effective managers. Ironically, the managers who could gain the most from valid developmental feedback may be less likely to receive it.

Because the level of rater agreement could influence the effectiveness of MSF, focal managers should be informed of the rater group's level of agreement during the feedback process. Most multisource feedback instruments provide some indication of the variance of a source's ratings (London & Smither, 1995). For example, the Benchmarks® feedback report provides the standard deviations of peer and subordinate ratings for each rating dimension. However, in addition to providing a manager with the agreement or variation of the raters, some guidelines for interpreting the level of agreement among raters should also be given. First, focal managers should be given norms for what level of agreement is typical. Second, focal managers should be coached on how to use information about agreement for their development. For example, feedback sessions often ask focal managers to examine areas where there are discrepancies *among* rating sources, which can highlight potential developmental needs. Similarly, having focal managers examine areas where there is relatively low agreement *within* a source could suggest some targeted action aimed at a specific rating source. For instance, if there is low agreement among subordinates about how much development the focal manager is providing, the focal manager may want to reflect on the extent to which he or she is developmentally stretching each direct report. It is possible that this type of discrepancy may uncover the need to provide better development to his or her entire team. Thus, information about the extent of agreement should be considered along with the average rating score to help the focal manager triangulate areas for development. Providing the level of agreement will present more targeted and precise information to the focal manager. By including this information, focal managers can formulate more nuanced personal development plans that integrate feedback regarding the magnitude of ratings as

well as the variance of these ratings. Thus, incorporating rater agreement into feedback may complicate the process, but the end result may be well worth it.

Another practical implication of the present research concerns the opportunity that different rating constituents have to view the focal manager performing all of the managerial behaviors that are captured in a typical MSF instrument. The results of the pretest suggest that peers and subordinates have different opportunities to observe various types of managerial behaviors. Being able to observe these behavioral domains is likely to pose the greatest problem when a rater is not well acquainted with the focal manager. In such situations, the ratings will most likely be of poor quality. The ratings may be based more on hearsay or stereotypes than the focal manager's actual behavior.

As a solution, MSF instruments could be tailored so that rating groups only rate behavioral domains that are relevant to their relationship with the focal manager. For instance, based on the present study's pretest, raters felt that subordinates generally have more opportunities than peers to see the focal manager leading others. In addition, peers were rated to have more information about the focal manager's career management skills than subordinates. Additional research, using a larger sample of managers could confirm whether these findings generalize to a managerial population. The results of such a study could be used to create a more tailored MSF instrument. An alternate solution would be to have individual raters determine whether they have enough information to make specific types of ratings. A MSF instrument could be designed with a "not enough information" response option so that raters only assign a rating when they have enough information to make a valid rating. Both solutions could result in a more accurate

assessment of the focal manager by eliminating ratings that are based on best guesses rather than actual behavioral evidence.

Limitations and future research. One goal of the present research was to examine the level of agreement within MSF rating groups using the relatively new a_{wg}. Although there were some benefits to using this method for assessing agreement, there were also some inherent limitations. One important limitation was the inability to calculate agreement for rating groups whose mean ratings were outside of the interpretable range of values. Another more general limitation was the relative lack of variation in agreement across rating groups, which was problematic in testing my hypotheses.

Eliminating between 30 and 40% of rater groups because their ratings fell outside of the interpretable range begs the question of whether the sample of rater groups with a_{wg} indices is representative of the agreement of the larger sample. In particular, rating dimensions that had higher mean ratings also had more rater groups removed for a_{wg} calculations because their observed mean was more likely to fall outside of the interpretable range (e.g., the *Differences Matter* and *Being a Quick Study* dimensions). Despite this systematic lack of agreement data, I feel confident that the sample of rater groups with a_{wg} scores did not bias the findings of the present study. I compared a_{wg} with r_{wg} indices, which were calculated for all rater groups, regardless of group size or mean rating. First, r_{wg} indices were very highly correlated with a_{wg} indices. Also, I tested my hypotheses using both r_{wg} indices and found similar outcomes as analyses with a_{wg}. Therefore, the large proportion of rating groups who were removed from my sample should not be a concern for interpreting my research findings. Still, researchers who choose to assess agreement using a_{wg} in the future should weigh the benefits of this index

(e.g., lack of scale dependency) against some of the potential drawbacks such as narrow interpretable ranges for small rater groups.

A more general limitation of the present research was the skewed distribution of a_{wg}. Findings from prior MSF studies suggest that ratings tend to be negatively skewed (e.g., Mount, 1984). Moreover, the Benchmarks® instrument is often used for the development of managers who are relatively accomplished (LeBreton et al, 2003), and as such, the within-source agreement may have had more range restriction than it would have for less accomplished managers. The inherent limitation with the negatively skewed distribution of a_{wg} was that it limited the magnitude of relationships that I was able to find with the predictor variables. Correlation coefficients generally decrease as the range, and consequently, the standard deviation of a variable decreases (Cohen et al., 2003). Thus, one major limitation of the present research was in my ability to find meaningful relationships between my predictors and within-source agreement because of the lack of variation in within-source agreement.

In addition to the range restriction issue, prior research has demonstrated that analyses with a_{wg} may be especially prone to committing a Type II error (Roberson et al., 2007), or failing to find a significant effect when one exists. Comparing indices of agreement (i.e., r_{wg}, r^*_{wg}, *SD*, *AD*, *V*), Roberson et al. (2007) examined the probability for committing Type II errors using a Monte Carlo simulation. Generally, agreement indices had a high incidence of committing a Type II error when testing for a strength effect. A strength effect is similar to the design of the present research which tested the relationships between group-level constructs (e.g., acquaintance, diversity) and the group's level of agreement. Specifically, Roberson et al. found that, "true relationships

will be detected less than 30% of the time" (p. 584). Even more troubling was their finding that a_{wg} (and V) performed poorly relative to the other indices in terms of committing a Type II error. Taken together, Roberson et al.'s conclusions suggest that the power to detect relationships between a_{wg} and the predictors of the present research may have been quite low. Thus, even finding small relationships between the predictors and agreement, especially when these relationships were consistent across most of the 16 rating dimensions, should be interpreted as notable given the high level of range restriction for a_{wg} and a high probability of committing a Type II error. Therefore, it is possible that the strength of some relationships reported for the present research may have been obscured as a consequence of using the a_{wg} index.

One challenge in the present research was the difficulty in disentangling agreement that may have been result of response bias from agreement that accurately reflects the focal manager's behavior. Although job performance ratings may never be completely free from rating bias, the use of assessment center ratings could be helpful in disentangling these issues. Some prior research has demonstrated the validity of assessment center ratings (Kolk, Born, & van der Flier, 2002; Lievens, 2002). These ratings may also be less motivationally biased than MSF ratings because the assessors do not have a working relationship with focal managers. Thus, future research could compare assessment center ratings with MSF ratings to further investigate the complex relationship between rater agreement and accuracy.

Another limitation of the present research was that a few of the predictor variables were proxy variables that did not directly measure the constructs of interest. For example, the demographic diversity variables were proxies for the extent to which rater groups

would differently interpret and rate a focal rater's performance. Although prior research suggests that individuals who are demographically different are less likely to communicate with one another and interpret information differently (Rentsch & Klimoski, 2001; Zenger & Lawrence, 1989), it would be preferable to directly measure the diversity of a rater group in terms of how they interpret and rate performance. This type of research could be especially valuable because I found consistent, albeit very weak relationships between some demographic composition variables and within-source agreement. Measuring process variables may result in finding stronger relationships between deep-level diversity and within-source agreement, as has been shown by prior research (e.g., Harrison et al., 1998).

In addition, the conceptualization of demographic composition was limited in that it only accounted for a group's composition across a demographic variable and did not make specific predictions about how the composition of specific status characteristics might uniquely influence the level of agreement among groups, as some researchers recommend investigating (DiTomaso, Post, & Parks-Yancy, 2007; Hewstone et al., 2006; van Knippenberg & Schippers, 2007). For instance, it is possible that a homogenous group of women might differ from a homogenous group of men in their level of agreement with one another. By ignoring status characteristics associated with gender and race, the unique effects of these characteristics were not tested. Moreover, I chose to focus on the rater group's composition. However, it is possible that the composition of the rater groups may interact with the characteristics of the focal manager as has been discussed by other researchers (Chattopadhyay, Tluchowska, & George, 2004; Somech, 2003; Tsui, Egan, & Xin, 1995; Tsui & O'Reilly, 1989). For example, the level of

agreement among a group of women rating a female manager might systematically differ from women rating a male manager. Taken together, future research should be designed to further examine the relationship between demographic composition and within-source agreement by including 1) process effects and deep-level forms of diversity (e.g., leadership preferences) and 2) the specific effects of particular demographic characteristics.

Other process-related variables should be examined in the future as they relate to within-source agreement. For instance, I found a small, negative relationship between acquaintance and within-source agreement. One way to explain this contradictory finding is that raters who have low levels of acquaintance may make biased ratings, particularly for content areas where they have little firsthand knowledge of the behaviors being rated. To directly test this hypothesis, future research could be designed to measure the extent to which individual raters are able to observe specific types of behavior. In addition, being relatively inexperienced at rating may explain why subordinates have lower agreement (e.g., Bates, 2002). Future research could test this hypothesis by examining how organizational level and years of experience relate to agreement. By controlling for level and experience, we could learn whether these factors account for the differences in agreement that I found between peers and subordinates.

Another limitation of the present research was the use of MBTI to measure personality. The research hypotheses regarding extraversion, agreement, and conscientiousness were based on prior research on the Big 5 personality factors (e.g., Colvin, 1993) measured by the NEO PI-R (Costa & McCrae, 1992). The present research, in contrast, used the MBTI personality measures to test hypotheses. Although

research by Furnham et al, (2003) suggested that the MBTI factors relate to the Big 5 dimensions, some factors are only moderately correlated (i.e., agreeableness and conscientiousness). Thus, some relationships between personality traits and judgability may not have been fully tested in the present research. Future research should further explore the relationship between focal manager personality and agreement using the Big 5 personality traits, as well as other personality measures.

This study examined characteristics of the rater group, the focal manager, and rating dimensions that differentiate high levels of agreement from low levels of agreement. However, it is possible that there are other predictors of agreement that were not tested in the present research. For instance, the quality of subordinates' leader-member exchange relationships may predict within-source agreement (London & Wohlers, 1991). Similarity in the quality of LMX relationships among subordinates might relate to higher agreement as opposed to groups that have LMX relationships of varying quality. Other personality traits may also relate to within-source agreement. For example, a manager who is a high self-monitor may actively adjust his or her style to the particular rater. If this is the case, it is likely that agreement would be lower for a manager who is a high self-monitor compared to a low self-monitor.

Conclusion. With the widespread use of multisource feedback tools (Chappelow, 1994), it is important to understand some of the underlying assumptions that contribute to the effectiveness of multisource instruments. One assumption, the relatively high agreement of raters within rating sources (e.g., Borman, 1997), has largely been ignored by a majority of prior research on MSF. However, the present research suggests that understanding within-source agreement is important. Although I found indications that

rater groups generally agreed with one another, there was still a fairly large portion of rater groups who failed to reach acceptable levels of agreement. In addition, I found that higher levels of agreement were associated with more effective managers, suggesting that highly effective managers may be easier to rate than managers who have more variable track records. My findings underscore the importance of understanding the predictors of agreement among raters and how levels of agreement may influence the effectiveness of MSF instruments.

Table 1

Demographic Characteristics of Focal Managers

Characteristic		n	%
Gender			
	Female	10,770	32.0
	Male	22,873	67.9
	Missing	53	0.2
Race			
	African American	1,394	4.1
	American Indian or Alaskan Nat	129	0.4
	Asian or Pacific Islander	1,200	3.6
	Caucasian	27,644	82.0
	Hispanic	869	2.6
	Multiracial	252	0.7
	Other	1,040	3.1
	Missing	1,168	3.5
Highest education level completed			
	High school	2,421	7.2
	Associate's degree	1,256	3.7
	Bachelor’s degree	15,532	46.1
	Master's degree	11,073	32.9
	Doctorate or professional degree	2,775	8.2
	Missing	639	1.9
Organizational level			
	First level	707	2.1
	Middle	8,538	25.3
	Upper middle	13,882	41.2
	Executive	7,950	23.6
	Top management	1,287	3.8
	Missing	1,332	4.0
Level of experience in job			
	No experience	8,051	23.9
	Moderately experienced	18,672	55.4
	Very experienced	6,696	19.9
	Missing	277	0.8

Note: N = 33,696.

Table 2

The Number and Proportion of Peer Rater Groups Who Met Requirements For Calculating a_{wg}

	Total number of rater groups	Groups with four or more raters		Groups where calculations of a_{wg} were possible	
Benchmarks® Scale	n	n	%	n	%
Resourcefulness	32,356	17,742	55	10,981	34
Doing Whatever it Takes	32,351	17,639	55	11,737	36
Being a Quick Study	32,352	17,536	54	9,860	30
Decisiveness	32,352	17,676	55	12,129	37
Leading Employees	32,316	16,717	52	13,274	41
Confronting Employees	32,217	15,320	48	12,821	40
Building/Mending Relationships	32,356	17,724	55	12,563	39
Compassion and Sensitivity	32,268	15,959	49	11,656	36
Straightforwardness	32,357	17,831	55	12,223	38
Balance b/w Life & Work	32,318	16,894	52	10,644	33
Self-Awareness	32,340	17,326	54	13,387	41
Putting People at Ease	32,363	17,970	56	8,873	27
Differences Matter	32,328	17,060	53	8,744	27
Participative Management	32,339	17,474	54	12,951	40
Career Management	32,310	16,377	51	12,604	39
Change Management	32,325	17,155	53	13,128	41
M	32,328	17,150	53	11,723	36

Note: Percentages are based on the proportion of group compared to the total number of rater groups. Calculations of a_{wg} were possible if the rater group had a least four raters and the group mean was within the interpretable range of values.

Table 3

The Number and Proportion of Subordinate Rater Groups Who Met Requirements For Calculating a_{wg}

	Total number of rater groups	Groups with four or more raters		Groups where calculations of a_{wg} were possible	
Benchmarks® Scale	n	n	%	n	%
Resourcefulness	29,961	17,002	57	8,127	27
Doing Whatever it Takes	29,963	16,996	57	8,967	30
Being a Quick Study	29,953	16,824	56	7,955	27
Decisiveness	29,965	17,049	57	9,777	33
Leading Employees	29,958	17,054	57	11,793	39
Confronting Employees	29,840	16,083	54	12,421	42
Building/Mending Relationships	29,961	16,957	57	10,709	36
Compassion and Sensitivity	29,944	16,866	56	10,626	35
Straightforwardness	29,969	17,130	57	10,108	34
Balance b/w Life & Work	29,929	16,388	55	10,434	35
Self-Awareness	29,951	16,723	56	11,840	40
Putting People at Ease	29,975	17,203	57	8,332	28
Differences Matter	29,949	16,824	56	7,766	26
Participative Management	29,958	17,050	57	11,254	38
Career Management	29,925	16,323	55	11,330	38
Change Management	29,942	16,847	56	11,110	37
M	29,946	16,832	56	10,159	34

Note: Percentages are based on the proportion of group compared to the total number of rater groups. Calculations of a_{wg} were possible if the rater group had a least four raters and the group mean was within the interpretable range of values.

Table 4

Means, Standard Deviations, and Intercorrelations for Peer Within-Source Agreement (a_{wg}) and Predictor Variables

Variable	*M*	*SD*	*N*	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Resourcefulness a_{wg}	0.88	0.10	10,981	--												
2. Doing Whatever it Takes a_{wg}	0.87	0.11	11,737	.74(**)	--											
3. Being a Quick Study a_{wg}	0.84	0.15	9,860	.59(**)	.59(**)	--										
4. Decisiveness a_{wg}	0.85	0.13	12,129	.53(**)	.60(**)	.45(**)	--									
5. Leading Employees a_{wg}	0.88	0.11	13,274	.68(**)	.68(**)	.50(**)	.50(**)	--								
6. Confronting Employees a_{wg}	0.86	0.12	12,821	.53(**)	.54(**)	.37(**)	.47(**)	.60(**)	--							
7. Building/Mending Relationships a_{wg}	0.86	0.12	12,563	.70(**)	.62(**)	.48(**)	.46(**)	.72(**)	.47(**)	--						
8. Compassion and Sensitivity a_{wg}	0.87	0.11	11,656	.52(**)	.49(**)	.39(**)	.37(**)	.68(**)	.43(**)	.61(**)	--					
9. Straightforwardness a_{wg}	0.83	0.14	12,223	.57(**)	.51(**)	.40(**)	.42(**)	.57(**)	.40(**)	.66(**)	.49(**)	--				
10. Balance b/w Life & Work a_{wg}	0.82	0.16	10,644	.28(**)	.25(**)	.25(**)	.25(**)	.32(**)	.23(**)	.31(**)	.36(**)	.28(**)	--			
11. Self-Awareness a_{wg}	0.82	0.16	13,387	.58(**)	.55(**)	.45(**)	.40(**)	.63(**)	.43(**)	.67(**)	.54(**)	.56(**)	.30(**)	--		
12. Putting People at Ease a_{wg}	0.77	0.19	8,873	.44(**)	.39(**)	.29(**)	.29(**)	.47(**)	.28(**)	.58(**)	.52(**)	.47(**)	.31(**)	.45(**)	--	
13. Differences Matter a_{wg}	0.83	0.15	8,744	.48(**)	.45(**)	.39(**)	.36(**)	.55(**)	.39(**)	.55(**)	.59(**)	.48(**)	.32(**)	.49(**)	.43(**)	--
14. Participative Management a_{wg}	0.86	0.12	12,951	.64(**)	.58(**)	.46(**)	.47(**)	.73(**)	.47(**)	.75(**)	.65(**)	.57(**)	.31(**)	.64(**)	.48(**)	.58(**)
15. Career Management a_{wg}	0.87	0.11	12,604	.62(**)	.60(**)	.44(**)	.44(**)	.68(**)	.47(**)	.66(**)	.61(**)	.51(**)	.31(**)	.62(**)	.45(**)	.53(**)
16. Change Management a_{wg}	0.88	0.11	13,128	.71(**)	.71(**)	.53(**)	.56(**)	.74(**)	.56(**)	.73(**)	.63(**)	.60(**)	.31(**)	.61(**)	.45(**)	.58(**)
17. Average of 16 Dimensions a_{wg}	0.88	0.11	20,181	.81(**)	.79(**)	.70(**)	.69(**)	.85(**)	.69(**)	.85(**)	.77(**)	.75(**)	.58(**)	.79(**)	.69(**)	.73(**)
18. Average Acquaintance	3.13	0.36	32,008	-.04(**)	-0.01	0.00	0.00	-.03(**)	-.05(**)	-.02(*)	-.03(**)	-.02(*)	-.06(**)	-.05(**)	0.00	-.03(**)
19. Gender Diversity	0.22	0.17	13,572	-.03(**)	-.04(**)	-.05(**)	-.03(**)	-.04(**)	-0.02	-.03(**)	-.04(**)	-.05(**)	-.05(**)	-.03(**)	-.05(**)	-.05(**)
20. Racial Diversity	0.13	0.16	12,743	-.04(**)	-.04(**)	-.04(**)	-.03(**)	-.04(**)	-.02(*)	-.05(**)	-.05(**)	-.04(**)	-.05(**)	-.02(*)	-.05(**)	-.08(**)
21. Age Diversity	0.15	0.07	12,732	-0.02	-.05(**)	-.04(**)	-.02(*)	-.05(**)	-.03(**)	-.04(**)	-.05(**)	-.03(**)	-.03(**)	-.04(**)	-0.02	-.03(*)
22. Educational Diversity	0.24	0.16	13,482	-.03(**)	-.03(**)	-0.01	-0.01	-.03(**)	-.03(**)	-.03(**)	-.02(*)	-.04(**)	-0.01	-.02(*)	-.03(*)	0.00
23. Extrovert/Introvert (MBTI)	-3.66	16.05	6,636	-0.02	0.00	-0.01	-0.03	-0.02	0.00	-0.03	-0.02	-.04(*)	-0.02	0.01	-0.01	-.05(*)
24. Sensing/iNtuition (MBTI)	-1.76	14.01	6,636	-0.01	-.06(**)	-0.02	-0.03	-0.03	0.00	-0.02	-0.02	0.00	-0.02	-.05(**)	-.05(*)	-0.01
25. Thinking/Feeling (MBTI)	-9.42	12.39	6,636	-0.03	-.06(**)	-0.03	-.04(*)	-.07(**)	-.07(**)	-0.04	-0.03	-0.01	-0.02	-0.04	-0.04	0.00
26. Judging/Perceiving (MBTI)	-5.98	15.30	6,636	-0.01	-0.01	0.00	-0.03	-0.01	-0.01	0.01	-0.02	0.00	-0.01	-0.02	-0.02	0.00
27. Total Inclusion (FIRO)	7.88	4.82	6,638	.06(**)	0.02	0.02	0.01	.05(**)	0.02	0.03	0.03	0.02	-0.01	0.00	0.03	0.01
28. Total Control (FIRO)	7.43	3.21	6,638	0.02	0.01	0.01	0.01	0.02	0.00	0.02	0.01	0.00	-0.01	-0.01	0.02	-0.02
29. Total Affection (FIRO)	9.60	4.03	6,638	0.00	-0.03	-0.02	0.00	0.01	0.00	0.00	0.01	-0.02	-0.01	-0.03	0.00	0.00
30. Overall Managerial Effectiveness	4.00	0.78	30,083	.09(**)	.10(**)	.08(**)	.07(**)	.07(**)	.06(**)	.08(**)	.03(**)	.06(**)	0.00	.11(**)	.04(**)	.05(**)

* $p < .05$. ** $p < .01$.

Table 4 continued

Means, Standard Deviations, and Intercorrelations for Peer Within-Source Agreement (a_{wg}) and Predictor Variables

Variable	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1. Resourcefulness a_{wg}																
2. Doing Whatever it Takes a_{wg}																
3. Being a Quick Study a_{wg}																
4. Decisiveness a_{wg}																
5. Leading Employees a_{wg}																
6. Confronting Employees a_{wg}																
7. Building/Mending Relationships a_{wg}																
8. Compassion and Sensitivity a_{wg}																
9. Straightforwardness a_{wg}																
10. Balance b/w Life & Work a_{wg}																
11. Self-Awareness a_{wg}																
12. Putting People at Ease a_{wg}																
13. Differences Matter a_{wg}																
14. Participative Management a_{wg}	--															
15. Career Management a_{wg}	.70(**)	--														
16. Change Management a_{wg}	.78(**)	.72(**)	--													
17. Average of 16 Dimensions a_{wg}	.84(**)	.80(**)	.86(**)	--												
18. Average Acquaintance	-0.01	-.03(**)	-.02(*)	-.02(**)	--											
19. Gender Diversity	-.04(**)	-.03(**)	-.03(**)	-.04(**)	-.03(**)	--										
20. Racial Diversity	-.03(**)	-.04(**)	-.05(**)	-.06(**)	-.03(**)	.03(**)	--									
21. Age Diversity	-.04(**)	-.05(**)	-.05(**)	-.05(**)	-.02(*)	.04(**)	0.02	--								
22. Educational Diversity	-.03(**)	-.02(*)	-.03(**)	-.03(**)	-.02(**)	.02(**)	0.00	.11(**)	--							
23. Extrovert/Introvert (MBTI)	0.00	-0.01	-0.02	-0.01	-.07(**)	0.02	0.00	-0.02	.04(*)	--						
24. Sensing/iNtuition (MBTI)	-.06(**)	-0.01	-0.02	-.04(**)	-0.02	0.03	.05(**)	-.06(**)	-0.03	-.14(**)	--					
25. Thinking/Feeling (MBTI)	-0.02	-0.02	-0.03	-.05(**)	.05(**)	.08(**)	.05(**)	0.00	.04(*)	-.17(**)	.29(**)	--				
26. Judging/Perceiving (MBTI)	-0.01	-0.02	-0.02	-0.02	0.02	0.02	0.03	-0.03	0.03	-.08(**)	.42(**)	.24(**)	--			
27. Total Inclusion (FIRO)	0.02	0.01	0.03	0.00	0.01	-0.03	-0.02	-0.02	-.04(*)	-.47(**)	.10(**)	.14(**)	.04(**)	--		
28. Total Control (FIRO)	0.00	0.01	0.00	0.00	.03(**)	-0.01	-0.03	0.00	-0.02	-.15(**)	0.02	-.03(**)	-.04(**)	.26(**)	--	
29. Total Affection (FIRO)	-0.03	-0.01	-0.01	-0.02	.05(**)	0.01	0.01	0.01	-.04(*)	-.46(**)	.11(**)	.28(**)	.05(**)	.62(**)	.14(**)	--
30. Overall Managerial Effectiveness	.06(**)	.08(**)	.08(**)	.05(**)	.08(**)	0.00	-.03(**)	0.00	0.00	-0.01	-.04(**)	-.03(*)	-.03(*)	-0.01	-0.01	.03(*)

* $p < .05$. ** $p < .01$.

Table 5

Means, Standard Deviations, and Intercorrelations for Subordinate Within-Source Agreement (a_{wg}) and Predictor Variables

Variable	*M*	*SD*	*N*	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Resourcefulness a_{wg}	0.86	0.12	8,127	--												
2. Doing Whatever it Takes a_{wg}	0.84	0.13	8,967	.79(**)	--											
3. Being a Quick Study a_{wg}	0.81	0.17	7,955	.62(**)	.63(**)	--										
4. Decisiveness a_{wg}	0.81	0.16	9,777	.58(**)	.64(**)	.49(**)	--									
5. Leading Employees a_{wg}	0.82	0.15	11,793	.72(**)	.73(**)	.52(**)	.53(**)	--								
6. Confronting Employees a_{wg}	0.82	0.15	12,421	.59(**)	.60(**)	.43(**)	.50(**)	.60(**)	--							
7. Building/Mending Relationships a_{wg}	0.84	0.14	10,709	.77(**)	.72(**)	.54(**)	.53(**)	.77(**)	.56(**)	--						
8. Compassion and Sensitivity a_{wg}	0.82	0.15	10,626	.57(**)	.56(**)	.41(**)	.41(**)	.71(**)	.48(**)	.67(**)	--					
9. Straightforwardness a_{wg}	0.80	0.16	10,108	.60(**)	.56(**)	.44(**)	.47(**)	.60(**)	.45(**)	.66(**)	.52(**)	--				
10. Balance b/w Life & Work a_{wg}	0.77	0.18	10,434	.32(**)	.30(**)	.28(**)	.30(**)	.34(**)	.27(**)	.36(**)	.41(**)	.32(**)	--			
11. Self-Awareness a_{wg}	0.78	0.19	11,840	.63(**)	.60(**)	.46(**)	.46(**)	.69(**)	.51(**)	.70(**)	.60(**)	.58(**)	.34(**)	--		
12. Putting People at Ease a_{wg}	0.73	0.21	8,332	.49(**)	.45(**)	.34(**)	.34(**)	.57(**)	.37(**)	.62(**)	.61(**)	.51(**)	.35(**)	.52(**)	--	
13. Differences Matter a_{wg}	0.78	0.19	7,766	.57(**)	.53(**)	.45(**)	.43(**)	.62(**)	.47(**)	.63(**)	.63(**)	.53(**)	.37(**)	.55(**)	.51(**)	--
14. Participative Management a_{wg}	0.82	0.15	11,254	.68(**)	.64(**)	.48(**)	.52(**)	.79(**)	.53(**)	.76(**)	.71(**)	.58(**)	.36(**)	.67(**)	.56(**)	.64(**)
15. Career Management a_{wg}	0.83	0.14	11,330	.68(**)	.66(**)	.48(**)	.51(**)	.74(**)	.55(**)	.72(**)	.65(**)	.56(**)	.35(**)	.67(**)	.54(**)	.62(**)
16. Change Management a_{wg}	0.85	0.13	11,110	.76(**)	.76(**)	.57(**)	.61(**)	.79(**)	.63(**)	.77(**)	.68(**)	.62(**)	.36(**)	.66(**)	.53(**)	.66(**)
17. Average of 16 Dimensions a_{wg}	0.84	0.14	17,741	.84(**)	.83(**)	.71(**)	.72(**)	.87(**)	.72(**)	.88(**)	.80(**)	.76(**)	.60(**)	.81(**)	.73(**)	.77(**)
18. Average Acquaintance	3.14	0.36	29,658	-.03(**)	-.03(**)	.00	-.03(**)	-.02	-.04(**)	-.01	-.01	-.03(**)	-.03(**)	-.04(**)	.00	-.04(**)
19. Gender Diversity	0.23	0.17	13,507	-0.01	-0.02	-0.02	-0.01	-.03(*)	-.04(**)	-.01	-.02(*)	.00	-.04(**)	-.01	-.01	-.04(**)
20. Racial Diversity	0.15	0.16	12,779	-.05(**)	-.03(**)	-.03(*)	-0.02	-.06(**)	-.05(**)	-.05(**)	-.05(**)	-.05(**)	-.05(**)	-.05(**)	-.05(**)	-.08(**)
21. Age Diversity	0.18	0.08	12,695	-.05(**)	-.03(*)	-.05(**)	-.03(*)	-.06(**)	-.06(**)	-.06(**)	-.08(**)	-.04(**)	-.06(**)	-.06(**)	-.04(**)	-.06(**)
22. Educational Diversity	0.30	0.19	13,407	-.05(**)	-.04(**)	-.03(*)	-0.02	-.03(**)	-.05(**)	-.05(**)	-.05(**)	-.05(**)	-.02(*)	-.03(**)	-.03(*)	-.04(**)
23. Extrovert/Introvert (MBTI)	-3.66	16.05	6,636	-.03	.01	.00	.01	.00	.01	.00	-.03	.00	.02	.00	-.03	.01
24. Sensing/iNtuition (MBTI)	-1.76	14.01	6,636	-.01	.00	-.02	.02	.02	.02	.01	.03	-.01	.03	.01	.01	.03
25. Thinking/Feeling (MBTI)	-9.42	12.39	6,636	-.01	-.03	.00	-.02	-.01	-.03	.01	.00	-.03	.00	-.02	-.07(**)	-.03
26. Judging/Perceiving (MBTI)	-5.98	15.30	6,636	.01	-.02	-.02	.01	.02	-.02	.01	.04(*)	.00	.04	.02	.01	.01
27. Total Inclusion (FIRO)	7.88	4.82	6,638	.05	.03	.04	-.01	.00	-.01	.04(*)	.03	.01	.01	.01	.02	.00
28. Total Control (FIRO)	7.43	3.21	6,638	-.05(*)	-.03	-.06(*)	-.04	.00	-.05(**)	-.02	-.01	-.02	-.03	.00	.00	.00
29. Total Affection (FIRO)	9.60	4.03	6,638	.02	.02	.01	.01	.01	.01	.05(*)	.01	.02	-.02	.02	.00	-.03
30. Overall Managerial Effectiveness	4.00	0.78	30,083	.08(**)	.08(**)	.07(**)	.05(**)	.06(**)	.04(**)	.07(**)	.05(**)	.07(**)	.00	.06(**)	.05(**)	.06(**)

* $p < .05$. ** $p < .01$.

Table 5 continued

Means, Standard Deviations, and Intercorrelations for Subordinate Within-Source Agreement (a_{wg}) and Predictor Variables

Variable	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1. Resourcefulness a_{wg}																
2. Doing Whatever it Takes a_{wg}																
3. Being a Quick Study a_{wg}																
4. Decisiveness a_{wg}																
5. Leading Employees a_{wg}																
6. Confronting Employees a_{wg}																
7. Building/Mending Relationships a_{wg}																
8. Compassion and Sensitivity a_{wg}																
9. Straightforwardness a_{wg}																
10. Balance b/w Life & Work a_{wg}																
11. Self-Awareness a_{wg}																
12. Putting People at Ease a_{wg}																
13. Differences Matter a_{wg}																
14. Participative Management a_{wg}	--															
15. Career Management a_{wg}	.76(**)	--														
16. Change Management a_{wg}	.81(**)	.77(**)	--													
17. Average of 16 Dimensions a_{wg}	.86(**)	.84(**)	.89(**)	--												
18. Average Acquaintance	-0.01	-.03(**)	-.02(*)	-.03(**)	--											
19. Gender Diversity	-0.01	-0.02	-0.01	-.02(*)	0.01	--										
20. Racial Diversity	-.05(**)	-.04(**)	-.06(**)	-.06(**)	-.06(**)	.04(**)	--									
21. Age Diversity	-.06(**)	-.06(**)	-.05(**)	-.08(**)	-.05(**)	.03(**)	.02(*)	--								
22. Educational Diversity	-.05(**)	-.04(**)	-.05(**)	-.05(**)	-.02(*)	-0.01	-0.01	.12(**)	--							
23. Extrovert/Introvert (MBTI)	0.00	0.01	0.01	-0.01	-.10(**)	-0.01	-0.01	0.01	0.02	--						
24. Sensing/iNtuition (MBTI)	-0.01	0.03	0.03	0.01	0.00	0.02	0.02	-.05(*)	-.05(**)	-.14(**)	--					
25. Thinking/Feeling (MBTI)	-0.01	0.00	-0.01	0.02	.06(**)	0.02	0.02	0.00	.03(*)	-.17(**)	.29(**)	--				
26. Judging/Perceiving (MBTI)	0.00	0.03	0.01	0.03	0.02	0.00	0.03	-.05(*)	-0.01	-.08(**)	.42(**)	.24(**)	--			
27. Total Inclusion (FIRO)	0.02	0.02	0.01	0.00	0.00	0.02	0.01	0.01	-0.03	-.47(**)	.10(**)	.14(**)	.04(**)	--		
28. Total Control (FIRO)	0.01	-0.03	-0.03	-.04(**)	0.00	-0.02	-0.01	-0.02	-0.03	-.15(**)	0.02	-.03(**)	-.04(**)	.26(**)	--	
29. Total Affection (FIRO)	0.01	0.00	0.00	0.02	.06(**)	0.01	0.02	0.02	-0.02	-.46(**)	.11(**)	.28(**)	.05(**)	.62(**)	.14(**)	--
30. Overall Managerial Effectiveness	.05(**)	.07(**)	.06(**)	.03(**)	.04(**)	0.01	0.00	-0.01	-.02(**)	-0.01	-.04(**)	-.03(*)	-.03(*)	-0.01	-0.01	.03(*)

* $p < .05$. ** $p < .01$.

Table 6

Regression Analyses Predicting Peer Agreement with Focal Manager Demographic Characteristics

	Benchmarks® Dimension															
	Resour	Doing	Quick	Decisiv	Leadin	Confr	Build	Compa	Straigh	Balanc	SelfA	Putting	Differ	Partic	Career	Chang
Predictor Variables																
Age	-.06***	-.06***	-.04***	-.05***	-.04***	-.06***	-.05***	-.02*	-.05***	-.02	-.03**	-.01	-.04**	-.05***	-.05***	-.05***
Tenure	.06***	.06***	.04**	.04**	.05***	.02*	.07***	.05***	.03**	-.02	.05***	.03*	.04**	.06***	.05***	.05***
Gender	.03**	.03**	.05***	.00	.05***	.02*	.05***	.06***	.04***	.07***	.05***	.07***	.04***	.04***	.04***	.04***
Experience	.00	.00	.00	.00	.02	.01	-.01	.01	.02	-.02	.00	-.01	-.02	.00	.00	.01
Race	.00	.01	.00	-.02	-.01	-.02	-.01	.00	-.01	.00	.00	.01	.02	-.01	-.01	-.01
Org. level	-.02*	-.01	.00	-.01	.00	-.03**	.00	.03*	-.01**	-.01	-.03*	.03*	.00	.00	.01	-.01
N	9932	10656	8884	10990	12075	11681	11409	10599	11060	9662	12182	8053	7934	11756	11471	11922
Summary statistics																
R	.08	.07	.07	.07	.07	.07	.09	.08	.07	.07	.08	.08	.07	.08	.07	.07
R^2	.01	.01	.01	.01	.01	.00	.01	.01	.00	.01	.01	.01	.01	.01	.00	.01

Note: Entries are standardized beta weights;
Resour = *Resourcefulness*, Doing = *Doing Whatever it Takes*, Quick = *Being a Quick Study*, Decisiv = *Decisiveness*, Leadin = *Leading Employees*, Confr = *Confronting Employees*, Build = *Building/Mending Relationships*, Compa = *Compassion and Sensitivity*, Straigh = *Straightforwardness*, Balanc = *Balance Between Life and Work*, SelfA = *Self-Awareness*, Putting = *Putting People at Ease*, Differ = *Differences Matter*, Partic = *Participative Management*, Career = *Career Management*, Chang = *Change Management*.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 7

Regression Analyses Predicting Subordinate Agreement with Focal Manager Demographic Characteristics

	Benchmarks® Dimension															
	Resour	Doing	Quick	Decisiv	Leadin	Confr	Build	Compa	Straigh	Balanc	SelfA	Putting	Differ	Partic	Career	Chang
Predictor Variables																
Age	-.07***	-.08***	-.06***	-.07***	-.06***	-.08***	-.07***	-.05***	-.07***	-.01	-.06***	-.06***	-.06***	-.07***	-.07***	-.06***
Tenure	.05***	.06***	.05***	.03*	.05***	.04***	.06***	.05***	.04***	.01	.04***	.03*	.04**	.06***	.04**	.05***
Gender	.05***	.01	.04**	.02*	.05***	.01	.06***	.08***	.05***	.06***	.05***	.07***	.06***	.04***	.04***	.03**
Experience	.01	.02	.01	.01	.01	.03***	.01	.03**	.01	.00	.02	.00	.00	.01	.01	.02*
Race	.02*	.00	.03*	.01	.01	.01	.02*	.00	.02*	.01	.01	.01	.02	.00	.02	.01
Org. level	-.02	.02	-.03*	-.02*	-.04***	-.02*	.03**	.05***	.02	-.01	.01	.01	.00	.03**	.02*	.02*
N	7308	8096	7169	8865	10684	11258	9679	9591	9112	9439	10706	7502	6999	10185	10240	10051
Summary statistics																
R	.10	.08	.09	.08	.08	.08	.10	.11	.09	.06	.08	.09	.09	.08	.08	.07
R^2	.01	.01	.01	.01	.01	.01	.01	.01	.01	.00	.01	.01	.01	.01	.01	.01

Note: Entries are standardized beta weights;
Resour = *Resourcefulness*, Doing = *Doing Whatever it Takes*, Quick = *Being a Quick Study*, Decisiv = *Decisiveness*, Leadin = *Leading Employees*, Confr = *Confronting Employees*, Build = *Building/Mending Relationships*, Compa = *Compassion and Sensitivity*, Straigh = *Straightforwardness*, Balanc = *Balance Between Life and Work*, SelfA = *Self-Awareness*, Putting = *Putting People at Ease*, Differ = *Differences Matter*, Partic = *Participative Management*, Career = *Career Management*, Chang = *Change Management*.
* $p < .05$. ** $p < .01$. ***$p < .001$.

Table 8

Means, Standard Deviations, Skewness, and Kurtosis of Predictor Variables for Peers and Subordinates

	Peers				Subordinates			
Predictor Variable	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis
Average Acquaintance	3.13	0.36	0.00	1.21	3.14	0.36	-0.07	1.29
Age Diversity	0.15	0.07	0.78	1.53	0.18	0.08	0.75	1.26
Gender Diversity	0.22	0.17	0.07	-1.16	0.23	0.17	0.00	-1.16
Educational Diversity	0.24	0.16	0.79	0.31	0.30	0.19	0.31	-0.64
Racial Diversity	0.13	0.16	0.91	-0.30	0.15	0.16	0.64	-0.77
TypeEI (MBTI)	0.58	0.49	-0.31	-1.90	0.58	0.49	-0.31	-1.90
TypeJP (MBTI)	0.62	0.49	-0.47	-1.78	0.62	0.49	-0.47	-1.78
TypeSN (MBTI)	0.56	0.50	-0.24	-1.94	0.56	0.50	-0.24	-1.94
TypeTF (MBTI)	0.78	0.41	-1.35	-0.17	0.78	0.41	-1.35	-0.17
Extrovert/Introvert Total (MBTI)	-3.66	16.05	0.14	-0.84	-3.66	16.05	0.14	-0.84
Judging/Perceiving Total (MBTI)	-5.98	15.30	0.11	-0.91	-5.98	15.30	0.11	-0.91
Sensing/iNtuition Total (MBTI)	-1.76	14.01	0.11	-0.41	-1.76	14.01	0.11	-0.41
Thinking/Feeling Total (MBTI)	-9.42	12.39	0.35	-0.15	-9.42	12.39	0.35	-0.15
Total Affection (FIRO)	9.60	4.03	0.17	-0.52	9.60	4.03	0.17	-0.52
Total Control (FIRO)	7.43	3.21	0.11	-0.44	7.43	3.21	0.11	-0.44
Total Inclusion (FIRO)	4.20	2.16	0.08	-1.23	7.88	4.82	0.08	-1.23
Overall Effectiveness	4.00	0.78	-0.79	0.29	4.00	0.78	-0.79	0.29

Note: Ns ranged between 6,636 and 30,083.

Table 9

Means, Standard Deviations, Skewness, and Kurtosis of Peer Agreement Indices

	a_{wg}				r_{wg}^{-un}				r_{wg}^{-ms}			
Benchmarks® Scale	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis
Resourcefulness	0.88	0.10	-2.26	9.32	0.89	0.12	-2.79	12.16	0.76	0.22	-1.57	2.32
Doing Whatever it Takes	0.87	0.11	-1.87	5.71	0.88	0.13	-2.57	9.90	0.74	0.24	-1.41	1.60
Being a Quick Study	0.84	0.15	-1.77	5.11	0.85	0.16	-2.33	7.38	0.69	0.27	-1.10	0.42
Decisiveness	0.85	0.13	-1.69	4.79	0.84	0.16	-2.16	6.19	0.67	0.28	-1.04	0.18
Leading Employees	0.88	0.11	-1.83	5.21	0.87	0.14	-2.40	8.22	0.72	0.25	-1.29	1.03
Confronting Problem Employees	0.86	0.12	-1.77	4.82	0.84	0.17	-2.12	5.76	0.66	0.29	-0.98	-0.04
Building/Mending Relationships	0.86	0.12	-1.84	5.19	0.86	0.15	-2.39	7.78	0.71	0.26	-1.23	0.75
Compassion/Sensitivity	0.87	0.11	-1.82	5.43	0.87	0.14	-2.38	8.12	0.72	0.26	-1.27	0.97
Straightforwardness and Composure	0.83	0.14	-1.55	3.68	0.83	0.17	-2.10	5.61	0.65	0.29	-0.94	-0.12
Balance Between Personal and Work Life	0.82	0.16	-1.68	4.51	0.82	0.18	-2.10	5.41	0.64	0.29	-0.89	-0.24
Self-Awareness	0.82	0.16	-1.67	3.94	0.81	0.19	-1.95	4.37	0.63	0.31	-0.79	-0.51
Putting People at Ease	0.77	0.19	-1.36	2.93	0.80	0.20	-1.82	3.70	0.61	0.31	-0.70	-0.71
Differences Matter	0.83	0.15	-1.80	5.14	0.85	0.16	-2.40	7.68	0.69	0.27	-1.14	0.51
Participative Management	0.86	0.12	-1.90	6.34	0.86	0.15	-2.29	7.17	0.70	0.27	-1.18	0.62
Career Management	0.87	0.11	-1.86	5.98	0.86	0.14	-2.34	7.72	0.71	0.26	-1.22	0.81
Change Management	0.88	0.11	-2.02	7.05	0.87	0.13	-2.45	8.77	0.73	0.25	-1.34	1.26
M	0.85	0.13	-1.79	5.32	0.85	0.16	-2.29	7.25	0.69	0.27	-1.13	0.55

Note: Ns ranged between 8,744-13,387 for a_{wg} and between 30,324-31,074 for r_{wg} indices.

Table 10

Means, Standard Deviations, Skewness, and Kurtosis of Subordinate Agreement Indices

	a_{wg}				r_{wg}^{-un}				r_{wg}^{-ms}			
Benchmarks® Scale	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis	M	SD	Skewness	Kurtosis
Resourcefulness	0.86	0.12	-2.21	9.73	0.88	0.13	-2.72	10.91	0.75	0.24	-1.52	1.97
Doing Whatever it Takes	0.84	0.13	-2.01	7.19	0.87	0.14	-2.56	9.10	0.72	0.25	-1.33	1.14
Being a Quick Study	0.81	0.17	-1.86	6.29	0.84	0.17	-2.28	6.79	0.67	0.28	-1.02	0.18
Decisiveness	0.81	0.16	-1.66	4.47	0.83	0.17	-2.11	5.51	0.65	0.29	-0.94	-0.13
Leading Employees	0.82	0.15	-1.73	4.47	0.82	0.18	-2.05	5.07	0.64	0.30	-0.90	-0.27
Confronting Problem Employees	0.82	0.15	-1.58	3.76	0.81	0.19	-1.89	4.24	0.62	0.31	-0.77	-0.55
Building/Mending Relationships	0.84	0.14	-1.87	5.72	0.85	0.16	-2.28	6.71	0.69	0.28	-1.14	0.38
Compassion/Sensitivity	0.82	0.15	-1.59	3.71	0.83	0.17	-2.08	5.36	0.66	0.29	-0.95	-0.10
Straightforwardness and Composure	0.80	0.16	-1.58	4.14	0.82	0.18	-2.00	4.83	0.63	0.30	-0.85	-0.38
Balance Between Personal and Work Life	0.77	0.18	-1.26	2.05	0.79	0.21	-1.69	2.99	0.58	0.32	-0.59	-0.92
Self-Awareness	0.78	0.19	-1.55	3.57	0.78	0.21	-1.67	2.83	0.57	0.32	-0.56	-0.97
Putting People at Ease	0.73	0.21	-1.29	2.34	0.79	0.22	-1.72	2.96	0.59	0.33	-0.61	-0.92
Differences Matter	0.78	0.19	-1.63	3.92	0.83	0.18	-2.19	5.70	0.65	0.29	-0.98	-0.07
Participative Management	0.82	0.15	-1.62	3.63	0.83	0.18	-2.04	5.07	0.65	0.29	-0.92	-0.20
Career Management	0.83	0.14	-1.66	4.24	0.83	0.17	-2.10	5.72	0.66	0.28	-0.96	-0.03
Change Management	0.85	0.13	-1.84	5.16	0.85	0.15	-2.25	6.97	0.69	0.27	-1.12	0.45
M	0.81	0.16	-1.68	4.65	0.83	0.18	-2.10	5.67	0.65	0.29	-0.95	-0.03

Note: Ns ranged between 7,766-12,421 for a_{wg} and between 28,199-28,620 for r_{wg} indices.

Table 11

Proportion of Peers and Subordinates with High and Low Levels of Agreement

Benchmarks® Scale	Proportion of Groups with High Agreement ($a_{wg} \geq .80$)		Proportion of Groups with Low Agreement ($a_{wg} < .60$)	
	Peers	Subordinates	Peers	Subordinates
Resourcefulness	84.2	76.6	2.1	4.2
Doing Whatever it Takes	81.4	72.5	2.5	5.3
Being a Quick Study	70.3	62.5	6.0	9.3
Decisiveness	71.8	63.0	5.2	9.1
Leading Employees	82.0	65.5	2.5	8.5
Confronting Problem Employees	75.7	64.8	4.4	8.2
Building/Mending Relationships	76.3	70.0	4.3	6.5
Compassion/Sensitivity	80.2	65.0	2.9	8.3
Straightforwardness and Composure	67.5	60.3	6.5	10.5
Balance Between Personal and Work Life	65.5	52.4	8.0	15.0
Self-Awareness	65.8	55.0	8.7	14.9
Putting People at Ease	50.7	43.7	15.8	21.7
Differences Matter	68.1	55.0	7.0	14.3
Participative Management	76.5	64.9	3.9	8.7
Career Management	79.6	69.1	3.0	6.3
Change Management	82.8	72.8	2.5	5.2
M	73.6	63.3	5.3	9.7

Table 12

Correlations Between Agreement and the Observed Mean for Peers and Subordinates

	a_{wg}		r_{wg}^{-un}	
Benchmarks® Scale	Peers	Subordinates	Peers	Subordinates
Resourcefulness	0.10**	0.14**	0.25**	0.33**
Doing Whatever it Takes	0.09**	0.12**	0.24**	0.33**
Being a Quick Study	0.14**	0.12**	0.28**	0.34**
Decisiveness	0.15**	0.15**	0.27**	0.33**
Leading Employees	0.06**	0.12**	0.16**	0.30**
Confronting Problem Employees	0.05**	0.10**	0.15**	0.23**
Building/Mending Relationships	0.15**	0.13**	0.29**	0.32**
Compassion/Sensitivity	0.04**	0.09**	0.18**	0.31**
Straightforwardness and Composure	0.14**	0.12**	0.29**	0.34**
Balance Between Personal and Work Life	0.16**	0.10**	0.30**	0.34**
Self-Awareness	0.19**	0.14**	0.28**	0.31**
Putting People at Ease	0.06**	0.05**	0.40**	0.46**
Differences Matter	0.14**	0.16**	0.31**	0.41**
Participative Management	0.12**	0.12**	0.23**	0.30**
Career Management	0.02*	0.04**	0.16**	0.22**
Change Management	0.07**	0.06**	0.15**	0.21**
M	0.11	0.11	0.25	0.32

Note: Observed means below scale midpoint (3.0) were reflected. Ns ranged between 7,766-13,387 for a_{wg} and between 28,197-31,074 for r_{wg}^{-un}.
* $p < .05$. ** $p < .01$.

Table 13

Within-Source Agreement Differences Between Peer Groups with Low Acquaintance and Those with High Acquaintance

	Low Acquaintance		High Acquaintance				
Benchmarks® Scale	M	SD	M	SD	df	t	r
Resourcefulness	0.89	0.10	0.87	0.13	2,801	2.71**	0.05
Doing Whatever it Takes	0.88	0.11	0.87	0.12	2,991	0.55	0.01
Being a Quick Study	0.84	0.15	0.85	0.16	2,572	-1.85	0.04
Decisiveness	0.85	0.13	0.85	0.15	3,171	-0.86	0.02
Leading Employees	0.88	0.10	0.87	0.11	3,415	2.77**	0.05
Confronting Problem Employees	0.86	0.12	0.85	0.14	3,359	3.53***	0.06
Building/Mending Relationships	0.86	0.12	0.86	0.13	3,223	1.11	0.02
Compassion/Sensitivity	0.88	0.11	0.87	0.12	2,972	2.10*	0.04
Straightforwardness and Composure	0.84	0.14	0.84	0.15	3,261	0.79	0.01
Balance Between Personal and Work Life	0.84	0.15	0.82	0.18	2,801	3.51***	0.07
Self-Awareness	0.84	0.15	0.82	0.17	3,510	3.19**	0.05
Putting People at Ease	0.76	0.19	0.78	0.19	2,342	-2.35*	0.05
Differences Matter	0.84	0.14	0.83	0.17	2,188	0.76	0.02
Participative Management	0.86	0.12	0.86	0.13	3,350	-0.24	0.00
Career Management	0.87	0.11	0.86	0.12	3,260	2.16*	0.04
Change Management	0.88	0.11	0.88	0.12	3,382	0.74	0.01
M	0.85	0.13	0.85	0.14			

*p < .05. ** p < .01. ***p < .001.

Table 14

Within-Source Agreement Differences Between Subordinate Groups with Low Acquaintance and Those with High Acquaintance

	Low Acquaintance		High Acquaintance				
Benchmarks® Scale	M	SD	M	SD	df	t	r
Resourcefulness	0.86	0.13	0.84	0.14	1,646	1.91	0.05
Doing Whatever it Takes	0.85	0.13	0.83	0.15	1,793	2.46*	0.06
Being a Quick Study	0.81	0.18	0.82	0.17	1,631	-0.69	0.02
Decisiveness	0.82	0.15	0.80	0.17	1,965	1.93	0.04
Leading Employees	0.82	0.15	0.82	0.16	2,318	0.29	0.01
Confronting Problem Employees	0.83	0.15	0.81	0.14	2,394	2.41*	0.05
Building/Mending Relationships	0.83	0.15	0.83	0.15	2,131	0.15	0.00
Compassion/Sensitivity	0.82	0.15	0.82	0.16	2,102	-0.24	0.01
Straightforwardness and Composure	0.81	0.16	0.80	0.17	2,075	2.06*	0.05
Balance Between Personal and Work Life	0.79	0.18	0.77	0.19	2,078	2.03*	0.04
Self-Awareness	0.79	0.18	0.77	0.20	2,324	2.82**	0.06
Putting People at Ease	0.73	0.21	0.74	0.22	1,718	-1.11	0.03
Differences Matter	0.95	0.09	0.95	0.10	1,550	2.99**	0.08
Participative Management	0.79	0.17	0.76	0.20	2,246	-0.37	0.01
Career Management	0.81	0.16	0.82	0.16	2,185	1.93	0.04
Change Management	0.84	0.14	0.82	0.15	2,182	0.64	0.01
M	0.82	0.14	0.84	0.14			

*$\underline{p} < .05$. ** $\underline{p} < .01$.

Table 15

Main Effects and Interaction Effects of Peer Acquaintance and Focal Manager Gender on Within-Source Agreement

	ANOVA		
	Confronting	Balance	Self-Awareness
Variable	F (1, 3352)	F (1, 2796)	F (1, 3504)
Acquaintance Level (A)	11.88***	14.57***	10.29***
Focal Manager Gender (G)	0.18	26.67***	6.66**
A x G	0.33	2.74	0.31

Note. F ratios are Wilk's approximation of Fs. ANOVA = univariate analysis of variance; Confronting = *Confronting Problem Employees*; Balance = *Balance Between Personal and Work Life.*
** p < .01. ***p < .001.

Table 16

Main Effects and Interaction Effects of Subordinate Acquaintance and Focal Manager Gender on Within-Source Agreement

	ANOVA		
	Confronting	Balance	Self-Awareness
Variable	F (1, 2388)	F (1, 2072)	F (1, 2318)
Acquaintance Level (A)	7.14**	4.02*	6.94**
Focal Manager Gender (G)	0.42	0.64	1.01
A x G	1.35	0.13	0.03

Note. F ratios are Wilk's approximation of Fs. ANOVA = univariate analysis of variance; Confronting = *Confronting Problem Employees*; Balance = *Balance Between Personal and Work Life.*
*p < .05. ** p < .01.

Table 17

Regression Analyses Predicting Peer Agreement with Demographic Composition Variables

	Benchmarks® Scale															
	Resour	Doing	Quick	Decisiv	Leadin	Confr	Build	Compa	Straigh	Balanc	SelfA	Putting	Differ	Partic	Career	Chang
Predictor Variable																
Age Diversity	-.02	-.04***	-.04**	-.02	-.04***	-.03**	-.04**	-.05***	-.03**	-.02	-.03**	-.01	-.03**	-.03*	-.04***	-.05***
Gender Diversity	-.05***	-.05***	-.06***	-.05***	-.04***	-.02*	-.04***	-.05***	-.05***	-.05***	-.03*	-.05***	-.05***	-.04***	-.04**	-.03**
Racial Diversity	-.04***	-.04***	-.04**	-.03**	-.05***	-.02*	-.05***	-.05***	-.05***	-.05***	-.03*	-.06***	-.07***	-.03**	-.04***	-.05***
Education Diversity	-.02	-.01	-.00	.00	-.02	-.02	-.02	-.01	-.03*	-.01	-.01	-.02	-.01	-.02	-.01	-.02
N	7983	8478	6741	8434	9680	9234	9086	8388	8547	7466	9271	6195	6185	9337	9159	9457
Summary statistics																
R	.07	.08	.08	.06	.08	.05	.08	.08	.08	.07	.06	.08	.10	.07	.07	.08
R^2	.01	.01	.01	.00	.01	.00	.01	.01	.01	.01	.00	.01	.01	.01	.01	.01

Note: Entries are standardized beta weights;
Resour = *Resourcefulness*, Doing = *Doing Whatever it Takes*, Quick = *Being a Quick Study*, Decisiv = *Decisiveness*, Leadin = *Leading Employees*, Confr = *Confronting Employees*, Build = *Building/Mending Relationships*, Compa = *Compassion and Sensitivity*, Straigh = *Straightforwardness*, Balanc = *Balance Between Life and Work*, SelfA = *Self-Awareness*, Putting = *Putting People at Ease*, Differ = *Differences Matter*, Partic = *Participative Management*, Career = *Career Management*, Chang = *Change Management*.
* $p < .05$. ** $p < .01$. ***$p < .001$.

Table 18

Regression Analyses Predicting Subordinate Agreement with Demographic Composition Variables

	Benchmarks® Scale															
	Resour	Doing	Quick	Decisiv	Leadin	Confr	Build	Compa	Straigh	Balanc	SelfA	Putting	Differ	Partic	Career	Chang
Predictor Variable																
Age Diversity	-.05***	-.03*	-.05***	-.02*	-.06***	-.05***	-.05***	-.08***	-.04**	-.05***	-.06***	-.04**	-.06***	-.06***	-.06***	-.05***
Gender Diversity	-.01	-.02	-.02	-.01	-.02	-.03**	-.01	-.02	-.01	-.04**	-.01	-.00	-.04**	-.00	-.02	-.01
Racial Diversity	-.05***	-.03*	-.03*	-.02*	-.06***	-.05***	-.05***	-.05***	-.04***	-.04***	-.05***	-.05***	-.08***	-.06***	-.04**	-.06***
Education Diversity	-.05***	-.04**	-.03*	.01	-.02*	-.04***	-.04***	-.05***	-.05***	-.01	-.02*	-.02	-.03*	-.05***	-.04**	-.04***
N	6197	6786	5764	7201	9029	9417	8161	8038	7425	7784	8801	6202	5871	8538	8638	8410
Summary statistics																
R	.09	.06	.07	.04	.09	.09	.09	.11	.08	.08	.08	.07	.11	.10	.08	.09
R^2	.01	.00	.01	.00	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01

Note: Entries are standardized beta weights;
Resour = *Resourcefulness*, Doing = *Doing Whatever it Takes*, Quick = *Being a Quick Study*, Decisiv = *Decisiveness*, Leadin = *Leading Employees*, Confr = *Confronting Employees*, Build = *Building/Mending Relationships*, Compa = *Compassion and Sensitivity*, Straigh = *Straightforwardness*, Balanc = *Balance Between Life and Work*, SelfA = *Self-Awareness*, Putting = *Putting People at Ease*, Differ = *Differences Matter*, Partic = *Participative Management*, Career = *Career Management*, Chang = *Change Management*.
* $p < .05$. ** $p < .01$. ***$p < .001$.

Table 19

Within-Source Agreement Differences Between Peer Groups with High Gender Diversity and Those with Low Gender Diversity

	High Diversity		Low Diversity				
Benchmarks® Scale	M	SD	M	SD	df	t	r
Resourcefulness	0.89	0.10	0.88	0.10	4,794	3.28**	0.05
Doing Whatever it Takes	0.88	0.10	0.87	0.11	5,157	3.79***	0.05
Being a Quick Study	0.84	0.13	0.82	0.15	4,038	4.84***	0.08
Decisiveness	0.84	0.13	0.83	0.13	5,100	3.18**	0.04
Leading Employees	0.88	0.10	0.87	0.11	5,923	3.82***	0.05
Confronting Problem Employees	0.86	0.12	0.85	0.12	5,664	1.82	0.02
Building/Mending Relationships	0.86	0.12	0.85	0.12	5,556	3.39***	0.05
Compassion/Sensitivity	0.88	0.10	0.86	0.11	5,111	4.13***	0.06
Straightforwardness and Composure	0.83	0.13	0.82	0.14	5,164	4.36***	0.06
Balance Between Personal and Work Life	0.82	0.15	0.80	0.16	4,509	4.27***	0.06
Self-Awareness	0.82	0.15	0.81	0.16	5,681	3.27**	0.04
Putting People at Ease	0.77	0.17	0.75	0.19	3,752	3.34***	0.05
Differences Matter	0.83	0.14	0.82	0.15	3,705	3.45***	0.06
Participative Management	0.87	0.11	0.85	0.12	5,735	3.79***	0.05
Career Management	0.87	0.10	0.86	0.11	5,613	3.47***	0.05
Change Management	0.88	0.10	0.87	0.11	5,762	3.27**	0.04
M	0.85	0.12	0.84	0.13			

Note: High Diversity = less than or equal to .05; Low Diversity = greater than .41.
** p < .01. ***p < .001.

Table 20

Within-Source Agreement Differences Between Subordinate Groups with High Gender Diversity and Those with Low Gender Diversity

	High Diversity		Low Diversity				
Benchmarks® Scale	M	SD	M	SD	df	t	r
Resourcefulness	0.86	0.13	0.85	0.12	2,778	0.95	0.02
Doing Whatever it Takes	0.85	0.13	0.83	0.14	3,044	2.07*	0.04
Being a Quick Study	0.81	0.16	0.79	0.16	2,527	2.64**	0.05
Decisiveness	0.81	0.16	0.80	0.17	3,235	1.37	0.02
Leading Employees	0.82	0.15	0.81	0.15	4,118	2.61**	0.04
Confronting Problem Employees	0.83	0.15	0.81	0.15	4,270	3.53***	0.05
Building/Mending Relationships	0.84	0.15	0.83	0.14	3,706	1.76*	0.03
Compassion/Sensitivity	0.82	0.15	0.81	0.15	3,656	2.32*	0.04
Straightforwardness and Composure	0.79	0.16	0.80	0.16	3,376	-0.33	0.01
Balance Between Personal and Work Life	0.77	0.17	0.75	0.18	3,519	3.91***	0.07
Self-Awareness	0.77	0.19	0.76	0.19	3,999	1.51	0.02
Putting People at Ease	0.72	0.20	0.72	0.20	2,882	0.75	0.01
Differences Matter	0.79	0.18	0.76	0.18	2,611	2.81**	0.05
Participative Management	0.82	0.15	0.81	0.15	3,872	0.79	0.01
Career Management	0.84	0.14	0.83	0.14	3,958	1.88	0.03
Change Management	0.85	0.14	0.84	0.13	3,779	1.17	0.02
M	0.81	0.16	0.80	0.16			

Note: High Diversity = less than or equal to .06; Low Diversity = greater than .42.
*p < .05. ** p < .01. ***p < .001.

Table 21

Within-Source Agreement Differences Between Peer Groups with High Racial Diversity and Those with Low Racial Diversity

	High Diversity		Low Diversity				
Benchmarks® Scale	M	SD	M	SD	df	t	r
Resourcefulness	0.88	0.10	0.87	0.11	5,323	4.16**	0.06
Doing Whatever it Takes	0.87	0.10	0.86	0.11	6,545	4.37**	0.05
Being a Quick Study	0.83	0.14	0.82	0.14	6,597	3.00**	0.04
Decisiveness	0.84	0.13	0.83	0.13	6,814	4.27**	0.05
Leading Employees	0.88	0.10	0.87	0.11	6,024	4.55***	0.06
Confronting Problem Employees	0.85	0.12	0.85	0.12	6,687	1.68	0.02
Building/Mending Relationships	0.86	0.12	0.84	0.13	6,024	2.72***	0.03
Compassion/Sensitivity	0.87	0.11	0.86	0.11	4,474	5.24***	0.08
Straightforwardness and Composure	0.83	0.13	0.81	0.14	6,054	3.23***	0.04
Balance Between Personal and Work Life	0.81	0.15	0.79	0.16	6,969	3.95***	0.05
Self-Awareness	0.81	0.15	0.80	0.15	6,731	2.94*	0.04
Putting People at Ease	0.76	0.18	0.73	0.18	4,478	4.25***	0.06
Differences Matter	0.83	0.14	0.80	0.16	4,789	2.89***	0.04
Participative Management	0.86	0.12	0.85	0.13	5,698	2.89**	0.04
Career Management	0.87	0.11	0.86	0.11	6,686	2.32**	0.03
Change Management	0.88	0.10	0.87	0.11	6,141	3.91***	0.05
M	0.85	0.12	0.83	0.13			

Note: High Diversity = less than or equal to .13; Low Diversity = greater than .29.
*p < .05. ** p < .01. ***p < .001.

Table 22

Within-Source Agreement Differences Between Subordinate Groups with High Racial Diversity and Those with Low Racial Diversity

	High Diversity		Low Diversity				
Benchmarks® Scale	M	SD	M	SD	df	t	r
Resourcefulness	0.86	0.12	0.84	0.13	4,400	3.39**	0.05
Doing Whatever it Takes	0.84	0.13	0.83	0.13	4,813	2.34*	0.03
Being a Quick Study	0.80	0.16	0.78	0.16	4,088	2.71**	0.04
Decisiveness	0.80	0.16	0.79	0.15	5,085	1.74	0.02
Leading Employees	0.82	0.14	0.80	0.15	6,374	5.76***	0.07
Confronting Problem Employees	0.82	0.14	0.80	0.16	6,638	5.13***	0.06
Building/Mending Relationships	0.84	0.14	0.82	0.14	5,796	4.46***	0.06
Compassion/Sensitivity	0.82	0.14	0.80	0.16	5,703	4.13***	0.05
Straightforwardness and Composure	0.80	0.16	0.78	0.16	5,254	3.79***	0.05
Balance Between Personal and Work Life	0.77	0.17	0.75	0.17	5,501	4.12***	0.06
Self-Awareness	0.77	0.18	0.75	0.19	6,226	4.11***	0.05
Putting People at Ease	0.72	0.20	0.70	0.20	4,406	3.17**	0.05
Differences Matter	0.78	0.17	0.75	0.19	4,206	5.53***	0.08
Participative Management	0.82	0.15	0.80	0.15	6,090	4.47***	0.06
Career Management	0.84	0.13	0.82	0.13	6,121	3.04**	0.04
Change Management	0.85	0.13	0.83	0.14	5,975	5.23***	0.07
M	0.81	0.15	0.79	0.16			

Note: High Diversity = less than or equal to .15; Low Diversity = greater than .33.
*p < .05. ** p < .01. ***p < .001.

Table 23

Regression Analyses Predicting Peer Agreement with Personality Variables

	Benchmarks® Dimension															
	Resour	Doing	Quick	Decisiv	Leadin	Confr	Build	Compa	Straigh	Balanc	SelfA	Putting	Differ	Partic	Career	Chang
Predictor Variables																
EI	-.03	-.01	-.02	-.04	-.03	-.01	-.03	-.03	-.04	-.02	.01	-.02	-.05*	-.01	-.01	-.02
TF	-.03	-.06**	-.03	-.04	-.07***	-.08***	-.05*	-.03	-.02	-.03	-.03	-.04	-.01	-.03	-.02	-.03
JP	-.01	-.01	-.00	-.02	.00	-.01	-.02	-.02	-.00	-.01	-.02	-.01	-.00	-.01	-.02	-.02
N	2,281	2,426	1,961	2,472	2,849	2,763	2,692	2,476	2,544	2,262	2,810	1,913	1,810	2,758	2,679	2,790
Summary statistics																
R	.04	.06	.03	.06	.07	.08	.05	.05	.04	.03	.04	.05	.05	.03	.03	.04
R^2	.00	.00	.00	.00	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Note: Entries are standardized beta weights;
Resour = *Resourcefulness*, Doing = *Doing Whatever it Takes*, Quick = *Being a Quick Study*, Decisiv = *Decisiveness*, Leadin = *Leading Employees*, Confr = *Confronting Employees*, Build = *Building/Mending Relationships*, Compa = *Compassion and Sensitivity*, Straigh = *Straightforwardness*, Balanc = *Balance Between Life and Work*, SelfA = *Self-Awareness*, Putting = *Putting People at Ease*, Differ = *Differences Matter*, Partic = *Participative Management*, Career = *Career Management*, Chang = *Change Management*.
* $p < .05$. ** $p < .01$. ***$p < .001$.

Table 24

Regression Analyses Predicting Subordinate Agreement with Personality Variables

	Benchmarks® Dimension															
	Resour	Doing	Quick	Decisiv	Leadin	Confr	Build	Compa	Straigh	Balanc	SelfA	Putting	Differ	Partic	Career	Chang
Predictor Variables																
EI	-.03	-.00	.00	.00	.00	.01	.01	-.03	-.01	.02	.00	-.04	.01	.00	.01	.01
TF	-.02	-.02	.01	-.03	-.01	-.03	.01	-.02	-.04	-.01	-.03	-.09**	-.03	-.01	-.01	-.01
JP	.02	-.01	-.02	-.02	.02	-.01	.01	.05*	.01	.05*	.02	.03	.02	.00	.03	-.02
N	1,613	1,798	1,591	1,967	2,449	2,580	2,252	2,187	2,084	2,177	2,439	1,764	1,549	2,319	2,352	2,268
Summary statistics																
R	.04	.03	.01	.03	.02	.04	.01	.06	.04	.05	.03	.09	.04	.01	.03	.02
R^2	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00

Note: Entries are standardized beta weights;
Resour = *Resourcefulness*, Doing = *Doing Whatever it Takes*, Quick = *Being a Quick Study*, Decisiv = *Decisiveness*, Leadin = *Leading Employees*, Confr = *Confronting Employees*, Build = *Building/Mending Relationships*, Compa = *Compassion and Sensitivity*, Straigh = *Straightforwardness*, Balanc = *Balance Between Life and Work*, SelfA = *Self-Awareness*, Putting = *Putting People at Ease*, Differ = *Differences Matter*, Partic = *Participative Management*, Career = *Career Management*, Chang = *Change Management*.
* $p < .05$. ** $p < .01$.

Table 25

Exploratory Regression Analyses Predicting Peer Agreement with Personality Variables

	Benchmarks® Dimension															
	Resour	Doing	Quick	Decisiv	Leadin	Confr	Build	Compa	Straigh	Balanc	SelfA	Putting	Differ	Partic	Career	Chang
Predictor Variables																
Inclusion	.09***	.07*	.05	.03	.08**	.03	.04	.04	.06*	.01	.04	.06	.02	.06*	.03	.05*
Control	.01	.01	.01	.01	.01	-.01	.02	.00	-.01	-.01	-.02	.02	-.02	-.01	.00	-.01
Affection	-.05*	-.06*	-.04	-.02	-.04	-.02	-.02	-.01	-.05*	-.02	-.05*	-.04	-.01	-.06*	-.02	-.03
SN	-.02	-.06**	-.03	-.03	-.03	.00	-.03	-.03	.00	-.02	-.05**	-.05*	-.01	-.06**	-.01	-.02
N	2,281	2,426	1,961	2,472	2,849	2,763	2,692	2,476	2,544	2,262	2,810	1,913	1,810	2,758	2,679	2,790
Summary statistics																
R	.08	.08	.05	.04	.07	.03	.05	.04	.05	.03	.07	.07	.02	.08	.03	.04
R^2	.01	.01	.00	.00	.01	.00	.00	.00	.00	.00	.01	.01	.00	.01	.00	.00

Note: Entries are standardized beta weights;
Resour = *Resourcefulness*, Doing = *Doing Whatever it Takes*, Quick = *Being a Quick Study*, Decisiv = *Decisiveness*, Leadin = *Leading Employees*, Confr = *Confronting Employees*, Build = *Building/Mending Relationships*, Compa = *Compassion and Sensitivity*, Straigh = *Straightforwardness*, Balanc = *Balance Between Life and Work*, SelfA = *Self-Awareness*, Putting = *Putting People at Ease*, Differ = *Differences Matter*, Partic = *Participative Management*, Career = *Career Management*, Chang = *Change Management*.
* $p < .05$. ** $p < .01$. ***$p < .001$.

Table 26

Exploratory Regression Analyses Predicting Subordinate Agreement with Personality Variables

	Benchmarks® Dimension															
	Resour	Doing	Quick	Decisiv	Leadin	Confr	Build	Compa	Straigh	Balanc	SelfA	Putting	Differ	Partic	Career	Chang
Predictor Variables																
Inclusion	.07*	.03	.08*	-.01	.00	-.01	.03	.05	-.01	.04	.03	.03	.03	.02	.03	.02
Control	-.07**	-.04	-.08**	-.04	-.01	-.05*	-.03	-.02	-.02	-.04	-.01	-.01	.00	.01	-.04	-.04
Affection	-.01	.01	-.02	.02	.01	.02	.03	-.03	.03	-.04	.02	-.02	-.05	.00	-.02	-.01
SN	-.01	-.01	-.03	.02	.02	.02	.00	.03	-.01	.03	.01	.01	.03	-.01	.03	.03
N	1,613	1,798	1,591	1,967	2,449	2,580	2,252	2,187	2,084	2,177	2,439	1,764	1,549	2,319	2,352	2,268
Summary statistics																
R	.08	.05	.09	.05	.02	.06	.06	.05	.03	.06	.02	.03	.05	.02	.05	.04
R^2	.01	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Note: Entries are standardized beta weights;
Resour = *Resourcefulness*, Doing = *Doing Whatever it Takes*, Quick = *Being a Quick Study*, Decisiv = *Decisiveness*, Leadin = *Leading Employees*, Confr = *Confronting Employees*, Build = *Building/Mending Relationships*, Compa = *Compassion and Sensitivity*, Straigh = *Straightforwardness*, Balanc = *Balance Between Life and Work*, SelfA = *Self-Awareness*, Putting = *Putting People at Ease*, Differ = *Differences Matter*, Partic = *Participative Management*, Career = *Career Management*, Chang = *Change Management*.
* $p < .05$. ** $p < .01$.

Table 27

Summary of Multiple Regression Analyses Predicting Average Within-Source Agreement among Peers with Personality and Acquaintance

Predictor Variables	(1)	(2)	(3)
Step 1[a]			
EI	-0.02	-0.02	-0.08
TF	-0.05**	-0.05**	-0.07
JP	-0.01	-0.01	0.16
Step 2			
Mean Acquaintance		-.04*	-.04
Step 3			
EI x Acquaintance			0.06
TF x Acquaintance			0.02
JP x Acquaintance			-0.17
Summary statistics[b]			
Multiple R	0.05	0.06	0.06
Multiple R^2	0.00	0.00	0.00
Multiple R^2 change for last step	0.00	0.00	0.00
F change for last step	3.59*	4.93*	0.40

Note: [a] Entries are standardized beta weights from full models; n = 4,077;
[b] Entries are for full models except for those indicated as change statistics.
* $p < .05$. ** $p < .01$.

Table 28

Summary of Multiple Regression Analyses Predicting Average Within-Source Agreement among Subordinates with Personality and Acquaintance

Predictor Variables	(1)	(2)	(3)
Step 1[a]			
EI	-0.01	-0.01	0.01
TF	0.01	0.01	0.18
JP	0.02	0.02	-0.03
Step 2			
Mean Acquaintance		-.03	-.04
Step 3			
EI x Acquaintance			-0.02
TF x Acquaintance			-0.16
JP x Acquaintance			0.06
Summary statistics[b]			
Multiple R	0.03	0.04	0.04
Multiple R^2	0.00	0.00	0.00
Multiple R^2 change for Last step	0.00	0.00	0.00
F change for last step	1.11	2.87	0.27

Note: [a] Entries are standardized beta weights from full models; n = 3,616;
[b] Entries are for full models except for those indicated as change statistics.

Table 29

Overview of Findings by Research Hypothesis

Hypothesis	Finding	Details	Dependent Variable
H1: The mean level of acquaintance within rating groups will be positively related to within-source agreement.	Not supported	Highest magnitude r = -.06; trend was inconsistent with hypothesis	a_{wg}: 16 dimensions peers & subs.
H2: The diversity of the rating group will be negatively related to within-source agreement.	Not supported; significant effects were very small, although in the predicted direction	Very small significant effects Maximum $R=.11;R^2 = .01$	a_{wg}: 16 dimensions peers & subs.
H3a: Within-source agreement will be higher for peers than subordinates.	Supported	Peer $a_{wg} = .85$ Sub $a_{wg} = .81$	Average a_{wg} of peers contrasted with subs.
H3b: Rater-group source will moderate the relationship between demographic characteristics and within-source agreement.	N/A – I did not test because I did not find a relationship between diversity and agreement		a_{wg}: 16 dimensions (peers & subs. Will be used together)
H4: Within peer and subordinate groups, agreement will be higher for rating dimensions in which the rater group has a greater opportunity to observe the behavioral domain as opposed to dimensions with fewer opportunities to observe representative behaviors.	Not supported	Not significant, but directionally consistent for peers (.86 vs. .84) and subordinates (.83 vs. .81); low power to test effect	High opportunity to observe dimensions vs. Low opportunity to observe dimensions
H5a: The level of focal manager's extraversion, agreeableness and conscientiousness will be positively related to within-source agreement.	Not supported	Maximum $R=.09;R^2 = .01$; no consistent relationships between personality and agreement	a_{wg}: 16 dimensions peers & subs.

H5b: The relationship between within-source agreement for peers and subordinates and the focal manager's extraversion will be stronger for rating dimensions that are more contingent on extraverted behaviors.	N/A - I did not test because I did not find a relationship between extraversion and agreement		a_{wg}: 16 dimensions peers & subs.
H5c: The relationship between the focal manager's personality and within-source agreement will be moderated by the extent that peers and subordinates are acquainted with the target such that the relationship between personality traits and agreement will be stronger with lower average levels of acquaintance.	Not supported		a_{wg}: 16 dimensions peers & subs.
RQ1: Do other personality traits (SN, inclusion, control, and affection) of the focal manager relate to within-source agreement?	Not supported	Maximum $R=.08; R^2 = .01$; no consistent relationships between personality and agreement	a_{wg}: 16 dimensions peers & subs.
H6: The mean performance rating of the focal manager will be positively related to within-source agreement.	Supported	Significant, small correlations for all dimensions but W-L Balance (average of .05 for peers and .06 for subordinates)	a_{wg}: 16 dimensions peers & subs.
RQ2: Which of the proposed predictors will contribute the most to our understanding of within-source agreement for peers and subordinates?	N/A		

Appendix A

Pilot Study: Peer's and Subordinate's Opportunity to Observe 16 Benchmarks® Dimensions

Instructions: Mark how frequently you think a **manager is able to directly observe his/her peers** (i.e., other managers who are at the same organizational level) performing the following managerial behaviors. If you are uncertain about a particular type of behavior, try your best to mark an answer.

	Behavior	1= Very Infrequently	2 = Infrequently	3 = Sometimes	4 = Frequently	5 = Very frequently
1.	Has solid working relationships with higher management.	1	2	3	4	5
2.	Is a visionary able to excite other people to work hard.	1	2	3	4	5
3.	Quickly masters new technical knowledge needed to do the job.	1	2	3	4	5
4.	Is action-oriented.	1	2	3	4	5
5.	Is willing to delegate important tasks, not just things he/she doesn't want to do.	1	2	3	4	5
6.	Can deal effectively with resistant employees.	1	2	3	4	5
7.	Gains commitment from others before implementing changes.	1	2	3	4	5
8.	Is straightforward with individuals about consequences of an expected action or decision.	1	2	3	4	5
9.	Relates to all kinds of individuals tactfully from the shop floor to top executives.	1	2	3	4	5
10.	Is willing to help an employee with personal problems.	1	2	3	4	5
11.	Remains calm when crises occur.	1	2	3	4	5
12.	Does not let job demands cause family problems.	1	2	3	4	5
13.	Does an honest self-assessment.	1	2	3	4	5
14.	Has a warm personality that puts people at ease.	1	2	3	4	5
15.	Understands and respects cultural, religious, gender, and racial differences.	1	2	3	4	5
16.	Actively seeks opportunities to develop professional relationships with others.	1	2	3	4	5

Note: Benchmarks® items are copyrighted material and therefore, only one item per dimension is presented for demonstrative purposes. The actual pretest consisted of 40 items.

Instructions: Mark how frequently you **think employees are able to directly observe their supervisor** performing the following managerial behaviors. If you are uncertain about a particular type of behavior, try your best to mark an answer.

	Behavior	1= Very Infrequently	2 = Infrequently	3 = Sometimes	4 = Frequently	5 = Very frequently
1.	Has solid working relationships with higher management.	1	2	3	4	5
2.	Is a visionary able to excite other people to work hard.	1	2	3	4	5
3.	Quickly masters new technical knowledge needed to do the job.	1	2	3	4	5
4.	Is action-oriented.	1	2	3	4	5
5.	Is willing to delegate important tasks, not just things he/she doesn't want to do.	1	2	3	4	5
6.	Can deal effectively with resistant employees.	1	2	3	4	5
7.	Gains commitment from others before implementing changes.	1	2	3	4	5
8.	Is straightforward with individuals about consequences of an expected action or decision.	1	2	3	4	5
9.	Relates to all kinds of individuals tactfully from the shop floor to top executives.	1	2	3	4	5
10.	Is willing to help an employee with personal problems.	1	2	3	4	5
11.	Remains calm when crises occur.	1	2	3	4	5
12.	Does not let job demands cause family problems.	1	2	3	4	5
13.	Does an honest self-assessment.	1	2	3	4	5
14.	Has a warm personality that puts people at ease.	1	2	3	4	5
15.	Understands and respects cultural, religious, gender, and racial differences.	1	2	3	4	5
16.	Actively seeks opportunities to develop professional relationships with others.	1	2	3	4	5

Note: Benchmarks® items are copyrighted material and therefore, only one item per dimension is presented for demonstrative purposes. The actual pretest consisted of 40 items.

Appendix B

Benchmarks® Section 1 Scales and Sample Items

Benchmarks® Scale	Sample Item
1) Resourcefulness (10 items)	▪ Has solid working relationships with higher management.
2) Doing Whatever it Takes (9 items)	▪ Is a visionary able to excite other people to work hard.
3) Being a Quick Study (4 items)	▪ Quickly masters new technical knowledge needed to do the job.
4) Decisiveness (4 items)	▪ Is action-oriented.
5) Leading Employees (14 items)	▪ Is willing to delegate important tasks, not just things he/she doesn't want to do.
6) Confronting Problem Employees (6 items)	▪ Can deal effectively with resistant employees.
7) Participative Management (10 items)	▪ Gains commitment from others before implementing changes.
8) Change Management (9 items)	▪ Is straightforward with individuals about consequences of an expected action or decision.
9) Building/Mending Relationships (11 items)	▪ Relates to all kinds of individuals tactfully from the shop floor to top executives.
10) Compassion and Sensitivity (7 items)	▪ Is willing to help an employee with personal problems.
11) Straightforwardness and Composure (4 items)	▪ Remains calm when crises occur.
12) Balance Between Life and Work (4 items)	▪ Does not let job demands cause family problems.
13) Self-Awareness (4 items)	▪ Does an honest self-assessment.
14) Putting People At Ease (4 items)	▪ Has a warm personality that puts people at ease.
15) Differences Matter (6 items)	▪ Understands and respects cultural, religious, gender, and racial differences.
16) Career Management (9 items)	▪ Actively seeks opportunities to develop professional relationships with others.

Appendix C

Lower and Upper Boundaries for Calculating a_{wg} by Group Size

Group Size	Lower boundary	Upper boundary
4	2.00	4.00
5	1.80	4.20
6	1.67	4.33
7	1.57	4.43
8	1.50	4.50
9	1.44	4.56
10	1.40	4.60
11	1.36	4.64
12	1.33	4.67
13	1.31	4.69
14	1.29	4.71
15	1.27	4.73
16	1.25	4.75
17	1.24	4.77
18	1.22	4.78
19	1.21	4.79
20	1.20	4.80
21	1.19	4.81
22	1.18	4.82
23	1.17	4.83
24	1.17	4.83
25	1.16	4.84

Note: Groups whose mean ratings fell outside of this range were removed from the sample with the exception of those whose mean equaled 5.0 or had perfect agreement.

References

Alliger, G. M., & Williams, K. J. (1989). Confounding among measures of leniency and halo. *Educational & Psychological Measurement, 49*, 1-10.

Allison, P. D. (1978). Measures of inequality. *American Sociological Review, 43*, 865-880.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

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

Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 63*, 975-985.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Bates, R. (2002). Liking and similarity as predictors of multi-source ratings. *Personnel Review, 31*, 540-552.

Baumeister, R. F., & Tice, D. M. (1988). Metatraits. *Journal of Personality, 56*, 571-598.

Beaty, J. C., Cleveland, J., & Murphy, K. R. (2001). The relation between personality and contextual performance in "strong" versus "weak" situations. *Human Performance, 14*, 125-148.

Becker, B. E., Huselid, M. A., & Ulrich, D. (2001). *The HR scorecard: Linking people strategy, and performance*. Cambridge, MA: Harvard Business Scool Press.

Becker, T. E., & Cote, J. A. (1994). Additive and multiplicative method effects in applied psychological research: An empirical assessment of three models. *Journal of Management, 20*, 625-641.

Bedeian, A. G., & Mossholder, K. W. (2000). The use of the coefficient of variation as a measure of diversity. *Organizational Research Methods, 3*, 285-297.

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

Borman, W. C. (1983). Implications of personality theory and research for the rating of work performance in organizations. In J. Cleveland (Ed.), *Performance measurement and theory*. Hillsdale, NJ: Erlbaum.

Borman, W. C. (1987). Personal construct, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an army sample. *Organizational Behavior & Human Decision Processes, 40*, 307-322.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In L. Hough (Ed.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. II). Palo Alto, CA: Consulting Psychologists Press.

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

Bozeman, D., P. (1997). Interrater agreement in multi-source performance appraisal: a commentary. *Journal of Organizational Behavior, 18*, 313-316.

Briggs, K. C., Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). Review of the Myers-Briggs Type Indicator Form M. In *the fourteenth mental*

measurements yearbook: Retrieved February 21, 2006, from the EBSCOhost Mental Measurements database.

Brown, R. D., & Hauenstein, N. M. A. (2005). Interrater Agreement Reconsidered: An Alternative to the r-sub(wg) Indices. *Organizational Research Methods, 8*, 165-184.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

Brutus, S., Fleenor, J. W., & London, M. (1998). Does 360-degree feedback work in different industries? A between-industry comparison of the reliability and validity of multi-source performance ratings. *Journal of Management Development, 17*, 177-190.

Brutus, S., Fleenor, J. W., & McCauley, C. D. (1999). Demographic and personality predictors of congruence in multi-source ratings. *Journal of Managerial Development, 18*, 417-435.

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5*, 159-172.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2*, 49-68.

Byrne, D., London, O., & Griffitt, W. (1968). The effect of topic importance and attitude similarity-dissimilarity on attraction in an intrastranger design. *Psychonomic Science, 11*, 303-304.

Byrne, D., London, O., & Reeves, K. (1968). The effects of physical attractiveness, sex, and attitude similarity on interpersonal attraction. *Journal of Personality, 36*, 259-271.

Byrne, D. E. (1971). *The Attraction Paradigm*. New York: Academic Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

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

Center for Creative Leadership. (2004). *Benchmarks facilitator's manual*. Greensboro, NC: Author.

Chan, D. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*, 300-311.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83*, 234-246.

Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality, 59*, 143-178.

Chaplin, W. F., & Panter, A. T. (1993). Shared meaning and the convergence among observers' personality descriptions. *Journal of Personality, 61*, 553-585.

Chappelow, C. T. (2004). 360- Degree Feedback. In E. Van Velsor (Ed.), *The Center for Creative Leadership handbook of leadership development* (2nd edition ed., pp. 58- 84). San Francisco: Jossey Bass.

Chattopadhyay, P., Tluchowska, M., & George, E. (2004). Identifying the ingroup: A closer look at the influence of demographic dissimilarity on employee social identity. *Academy of Management Review, 29*, 180-202.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement, 20*, 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Colvin, C. R. (1993). "Judgable" People: Personality, Behavior, and Competing Explanations. *Journal of Personality and Social Psychology, 64*.

Conway, J. M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management, 22*, 139-162.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric Properties of Multisource Performance Ratings: A meta-Analysis of Subordinate, Supervisor, Peer, and Self-Ratings. *Human Performance, 10*, 331-361.

Conway, J. M., Lombardo, K., & Sanders, K. C. (2001). A Meta-Analysis of Incremental Validity and Nomological Networks for Subordinate and Peer Rating. *Human Performance, 14*, 267-303.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory and five-factor inventory professional manual*. Odessa. FL: Psychological Assessment Resources.

Cox, T. H., Lobel, S. A., & McLeod, P. L. (1991). Effects of ethnic group cultural differences on cooperative and competitive behavior on a group task. *Academy of Management Journal, 34*, 827-847.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanich College Publishers.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity ". *Psychological Bulletin, 52*, 177-193.

Cronbach, L. J., Gleser, G. C., Nanada, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: John Wiley.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694.). Washington, D.C.: American Council on Education.

Curtis, A. B., Harvey, R. D., & Ravden, D. (2005). Sources of political distortions in performance appraisals: Appraisal purpose and accountability. *Group and Organization Management, 30*, 42-60.

Dansereau, F., Graen, G. B., & Hage, W. J. (1975). A vertical dyad linkage approach to leadership. *Organizational Behavior & Human Decision Processes, 13*, 46-78.

Deal, J. (2005). Generation gap: Fact or fiction? Retrieved February 9, 2005, http://www.ccl.org/CCLCommerce/news/newsletters/enewsletter/2005/FEBgap.aspx

Deal, J. J., & Stevenson, M. A. (1998). Perceptions of female and male managers in the 1990s: Plus ca change... *Sex Roles, 38*, 287-300.

DeShon, R. (2002). Generalizability Theory. In N. Schmitt (Ed.), *Measuring and analyzing behavior in organizations* (pp. 189-220). San Francisco: Jossey-Bass.

DiTomaso, N., Post, C., & Parks-Yancy, R. (2007). Workforce Diversity and Inequality: Power, Status, and Numbers. *Annual Review of Sociology, 33*, 473-501.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for r_{wg} and average deviation interrater agreement indexes. *Journal of Applied Psychology, 88*, 356-362.

Dunnette, M. D. (1966). *Personnel selection and placement*. Belmont, CA: Brooks-Cole.

Epitropaki, O., & Martin, R. (2004). Implicit Leadership Theories in Applied Settings: Factor Structure, Generalizability, and Stability Over Time. *Journal of Applied Psychology, 89*.

Feinberg, B. J., Ostroff, C., & Burke, W. W. (2005). The role of within-group agreement in understanding transformational leadership. *Journal of Occupational & Organizational Psychology, 78*, 471 - 488.

Field, A. (2005). *Discovering statistics through SPSS* (2nd ed.). Thousand Oaks, CA: Sage Publications Inc.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational & Psychological Measurement, 30*, 71-76.

Fleenor, J. W., Fleenor, J. B., & Grossnickle, W. F. (1996). Interrater reliability and agreement of performance ratings: a methodological comparison. *Journal of Business and Psychology, 10*, 367-380.

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

Fletcher, C. (1999). The implications of research on gender difference in self-assessment and 360 degree appraisal. *Human Resource Management Journal, 9*, 39-46.

Fletcher, C., & Baldry, C. (1999). Multi-source feedback systems: A research perspective. In I. T. Robertson (Ed.), *International review of industrial and organizational psychology 1999, Vol 14* (pp. 149-193). New York, NY: John Wiley & Sons Ltd.

Foti, R. J., & Lord, R. G. (1987). Prototypes and scripts: The effects of alternative methods of processing information on rating accuracy. *Organizational Behavior & Human Decision Processes, 39*, 318-340.

Freeberg, N. E. (1969). Relevance of rater-ratee acquaintance in the validity and reliability in ratings. *Journal of Applied Psychology, 53*, 518-524.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652-670.

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 55*, 149-158.

Funder, D. C., & Colvin, C. R. (1991). Explorations in Behavioral Consistency: Properties of Persons, Situations, and Behaviors. *Journal of Personality and Social Psychology, 60*.

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality & Social Psychology, 69*, 656-672.

Furnham, A., Moutafi, J., & Crump, J. (2003). The relationship between the revised NEO-Personality Inventory and the Myers-Briggs Type Indicator. *Social Behavior & Personality, 31*, 577-584.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Graen, G. B., & Uhl-Bien, M. (1995). Relationship based approach to leadership development of leader-member exchange (LMX) theory over 25 years: Applying a multilevel multidimensional perspective. *Leadership Quarterly, 6*, 219-247.

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

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

Greguras, G. J. R., Chet (1998). A New Look at Within-Source Interrater Reliability of 360-Degree Feedback Ratings. *Journal of Applied Psychology, 83*, 960-968.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

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

Harrison, D. A., Price, K. H., & Bell, M. P. (1998). Beyond relational demography: Time and the effects of surface- and deep-level diversity on work group cohesion. *Academy of Management Journal, 41*, 96-107.

Harter, J. K., Schmidt, F. L., & Hayes, T. L. (2002). Business-unit-level relationship between employee satisfaction, employee engagement, and business outcome: A meta-analysis. *Journal of Applied Psychology, 82*, 268-279.

Heilman, M. E. (1995). Sex stereotypes and their effects in the workplace: What we know and what we don't know. *Journal of Social Behavior and Personality, 10*, 3-26.

Heilman, M. E., Block, C. J., Martell, R. F., & Simon, M. C. (1989). Has anything changed? Current characterizations of men, women and managers, *Journal of Applied Psychology* (Vol. 74, pp. 935-942).

Hewstone, M., Crisp, R. J., Contarello, A., Voci, A., Conway, L., Marietta, G., et al. (2006). Tokens in the Tower: Perceptual Processes and Interaction Dynamics in Academic Settings with 'Skewed', 'Tilted' and 'Balanced' Sex Ratios. *Group Processes & Intergroup Relations, 9*, 509-532.

Hoffman, C. C. (1995). Applying range restriction corrections using published norms: Three case studies. *Personnel Psychology, 48*, 913-923.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Jackson, S. E., Brett, J. F., Sessa, V. I., Cooper, D. M., Julin, J. A., & Peyronnin, K. (1991). Some differences make a difference: Individual dissimilarity and group heterogeneity as correlates of recruitment, promotions, and turnover. *Journal of Applied Psychology, 76*.

Jackson, S. E., & Joshi, A. (2004). Diversity in social context: A multi-attribute, multilevel analysis of team diversity and sales performance. *Journal of Organizational Behavior, 25*, 675-702.

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

James, L. R., Demaree, R. G., & Wolf, G. (1993). Rwg: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78*, 306-309.

Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology, 86*, 984-996.

Johnson, J. W., & Ferstl, K. L. (1999). The effects of interrater and self-other agreement on performance improvement following upward feedback. *Personnel Psychology, 52*, 272-303.

Judge, T. A., & Ferris, G. R. (1993). Social Context of Performance Evaluation Decisions. *Academy of Management Journal, 36*, 80-106.

Jung, C. G. (1971). Psychological types (H. G. Baynes, Trans.). In R. F. C. Hull (Ed.), *The collected works of C. G. Jung* (Vol. 6th). Princeton: NJ: Princeton University Press.

Kaiser, R. B., & Craig, S. B. (2005). Building a better mouse trap: Item characteristics associated with discrepancies in 360-degree feedback. *Consulting Psychology Journal: Practice and Research, 57*, 235-245.

Kanter, R. M. (1977). Some effects of proportions on group life: Skewed sex ratios and responses to token women. *American Journal of Sociology, 82*, 954- 990.

Kavanaugh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin, 75*, 34-49.

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review, 98*, 155-163.

Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin, 102*.

Kenny, D. A., & Albright, L. (1994). Consensus in interpersonal perception: Acquaintance and the big five. 14p.

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist, 43*.

Kingstrom, P. O., & Mainstone, L. E. (1985). An investigation of the rater-ratee acquaintance and rater bias. *Academy of Management Journal, 28*, 641-653.

Kirchmeyer, C. (1995). Demographic similarity to the work group: A longitudinal study of managers at the early career stage. *Journal of Organizational Behavior, 16*, 67-83.

Klein, K., Conn, A. B., Smith, D. B., & Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-source agreement in employee perceptions of the work environment. *Journal of Applied Psychology, 86*, 3-16.

Klein, K., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review, 19*, 195-229.

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

Kolk, N. J., Born, M. P., & van der Flier, H. (2002). Impact of common rater variance on construct validity of assessment center dimension judgments. *Human Performance, 15*, 325-338.

Kozlowski, S., & Hattrup, K. (1992). A disagreement about within group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77*, 161-167.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology, 70*, 56-65.

Kraiger, K., & Teachout, M. S. (1990). Generalizability theory as construct-related evidence of the validity of job performance ratings. *Human Performance, 3*, 19-35.

Kurland, N. B., & Bailey, D. E. (1999). Telework: The advantage and challenges of working here, there, anywhere, and anytime. *Organizational Dynamics, 28*, 53-67.

Lance, C. E., & Teachout, M. S. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology, 77*.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 67*, 72-107.

Landy, F. J., & Guion, R. M. (1970). Development of scales for the measurement of work motivation. *Organizational Behavior & Human Performance, 5*, 93-103.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30*, 255-269.

Latham, G. P., & Wexley, K. N. (1982). *Increasing productivity through performance appraisal*. Reading, MA: Addison-Wesley.

Lawlis, G. F., & Lu, E. (1972). Judgment of counseling process: reliability, agreement, and error. *Psychological Bulletin, 78*, 17-20.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*, 80-128.

LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding r_{WG}, r^*_{WG}, $r_{(WG(J))}$, and $r^*_{(WG(J))}$. *Organizational Research Methods, 8*, 128-138.

LeBreton, J. M., & Senter, J. L. (in press). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*.

Lefkowitz, J. (1994). Sex-related job attitudes and dispositional variables: Now you see them,... *Academy of Management Journal, 37*, 323-349.

Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology, 87*, 675-686.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-Item ratings of a single target. *Applied Psychological Measurement, 23*, 127-135.

Lindsey, E. H., Homes, V., & McCall, M. W., Jr. (1987). *Key events in executives' lives*. Greensboro, NC: Center for Creative Leadership.

Lombardo, M., McCauley, C., McDonald, M. D., & Leslie, J. B. (2001). Review of Benchmarks revised. In *the fourteenth mental measurements yearbook [Electronic version]*: Retrieved February 21, 2006, from the EBSCOhost Mental Measurements database.

London, M., & Smither, J. W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance? Theory-based applications and directions for research. *Personnel Psychology, 48*, 803-839.

London, M., & Wohlers, A. J. (1991). Agreement between subordinate and self-ratings in upward feedback. *Personnel Psychology, 44*, 375- 390.

Lord, R. G. (1985). An information processing approach to social perceptions, leadership and behavioral measurement in organizations. *Research in Organizational Behavior, 7*, 87-128.

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*, 402-410.

Lord, R. G., Foti, R. J., & de Vader, C. L. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior & Human Performance, 34*, 343-378.

Lu, K. H. (1972). Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin, 78*, 17-20.

Malloy, T. E., Agatstein, F., Yarlas, A., & Albright, L. (1997). Effects of communication, information overlap, and behavioral consistency on consensus in social perception. *Journal of Personality and Social Psychology, 73*, 270-280.

McCauley, C. D., & Lombardo, M. M. (1990). Benchmarks: an instrument for diagnosing managerial strengths and weaknesses. In M. B. Clark (Ed.), *Measures of Leadership*. West Orange, NJ: Leadership Library of America.

McCauley, C. D., Lombardo, M. M., & Usher, C. J. (1989). Diagnosing management development needs: An instrument based on how managers develop. *Journal of Management, 15*, 389-403.

McCrae, R. R., & Costa, P. T. (1989). Reinterpreting the Myers-Briggs Type Indicator from the perspective of the five-factor model of personality. *Journal of Personality, 57*, 17-40.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*, 415-444.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality & Social Psychology, 90*, 862-877.

Mischel, W. (1969). Continuity and change in personality. *American Psychologist, 24*, 1012-1018.

Mischel, W. (1977). The interaction of person and situation. In N. S. Endler (Ed.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 333-352). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Mollica, K. A., & Treviño, L. K. (2003). Racial homophily and its persistence in newcomers' social networks. *Organization Science, 14*, 123-136.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics*. New York: McGraw Hill.

Mount, M. K. (1984). Psychometric properties of subordinate ratings of managerial performance. *Personnel Psychology, 37*, 687-701.

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.

Mount, M. K., Sytsma, M. R., Hazucha, J. F., & Holt, K. E. (1997). Rater-ratee race effects in developmental performance ratings of managers. *Personnel Psychology, 50*, 51-69.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873-900.

Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto: CA: Consulting Psychologists Press.

Nathan, B. R., & Alexander, R. A. (1983). The role of inferential accuracy in performance ratings. *Academy of Management Review, 10*, 109-115.

Offermann, L. R., Kennedy, J. K., & Wirtz, P. W. (1994). Implicit leadership theories: Content, structure, and generalizability. *Leadership Quarterly, 5*, 43-58.

O'Reilly, C. A., Caldwell, D. T., & Barnett, W. P. (1989). Work group demography, social integration, and turnover. *Administrative Science Quarterly, 34*, 21-37.

Organ, D. W. (1997). Organizational citizenship behavior: It's construct clean-up time. *Human Performance, 10*, 85-97.

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

Paulhus, D. L., & Bruce, M. N. (1992). The effect of acquaintanceship and validity in personality impressions: A longitudinal study. *Journal of Personality & Social Psychology, 63*, 816-824.

Paunonen, S. V. (1989). Consensus in personality judgments: Moderating effects of target-rater acquaintanceship and behavior observability. *Journal of Personality & Social Psychology, 56*, 823-833.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis*. Hillsdale, NJ: Lawrence Erlbaum Publishers.

Pfeffer, J. (1983). Organizational demography. *Research in Organizational Behavior, 5*, 299-357.

Phillips, J. S., & Lord, R. G. (1986). Notes on the practical and theoretical consequences of implicit leadership theories for the future of leadership measurement. *Journal of Management, 12*, 31-41.

Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of job performance rating: An examination of ratee race, ratee gender, and rater level effects. *Human Performance, 9*, 102-119.

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

Randel, A. E. (2002). Identity salience: A moderator of the relationship between group gender composition and work group conflict. *Journal of Organizational Behavior, 23*, 749-766.

Rentsch, J. R., & Klimoski, R. J. (2001). Why do "great minds' think alike?: Antecedents of team member schema agreement. *Journal of Organizational Behavior, 22*, 107-120.

Roberson, Q. M., Sturman, M. C., & Simons, T. L. (2006). Choosing between measures of climate strength: Much ado about nothing. *Manuscript submitted for publication.*

Roberson, Q. M., Sturman, M. C., & Simons, T. L. (2007). Does the measure of dispersion matter in multilevel research? A comparison of the relative performance of dispersion indexes. *Organizational Research Methods, 10*, 564-588.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Englewood Cliffs, NJ: Pearson/Prentice Hall.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75*, 322-327.

Sackett, P. R. (2000). Correction for Range Restriction: An Expanded Typology. *Journal of Applied Psychology, 85*, 112-118.

Sackett, P. R., & Ostgaard, D. J. (1994). Job-specific applicant pool and national norms for cognitive ability tests: Implications for range restriction correction in validation research. *Journal of Applied Psychology, 79*, 680-684.

Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology, 74*, 368-370.

Schmitt, M. J., & Ryan, A. M. (1997). Applicant withdrawl: The role of test-taking attitudes and racial differences. *Personnel Psychology, 50*, 855-877.

Schmitt, N., Noe, R. A., & Gottschalk, R. (1986). Using the lens model to magnify raters' consistency, matching, and shared bias. *Academy of Management Journal, 29*, 130-139.

Schneider, B. (1987). The people make the place. *Personnel Psychology, 40*, 437-453.

Schneider, B., Goldstein, H. W., & Smith, D. B. (1995). The ASA framework: An update. *Personnel Psychology, 48*, 747-779.

Schrader, B. W., & Steiner, D. D. (1996). Common comparison standards: An approach to Improving agreement between self and supervisory performance ratings. *Journal of Applied Psychology, 81*.

Schutz, W. C. (1957). *FIRO: The Three-Dimensional Theory*. New York: Rinehart & Company, Inc.

Schutz, W. C., Hammer, A. L., & Schnell, E. R. (2000). Review of the FIRO-B Fundamental Interpersonal Relations Orientation-Behavior. In *the fourteenth mental measurements yearbook*: Retrieved February 21, 2006, from the EBSCOhost Mental Measurements database.

Scullen, S. E., Mount, M. K., & Goff, M. I. (2000). Understanding the Latent Structure of Job Performance Ratings. *Journal of Applied Psychology, 85*, 956-970.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.

Shout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Smith, K. G., Smith, K. A., Olian, J. D., & Sims, H. P. (1994). Top management team demography and process: The role of social integration and communication. *Administrative Science Quarterly, 39*, 412-438.

Somech, A. (2003). Relationships of participative leadership with relational demography variables: A multi-level perspective. *Journal of Organizational Behavior, 24*, 1003-1018.

Sullivan, S. (1999). The changing nature of careers: A review and research agenda. *Journal of Management, 25*, 457-484.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*, 497-506.

Swann, W. B. (1984). Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review, 91*, 457-477.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of inter-group behavior. In L. W. Austin (Ed.), *Psychology of Intergroup Relations.* Chicago: Nelson-Hall.

Teachman, J. D. (1980). Analysis of population diversity. *Sociological Methods and Research, 8*, 341-362.

Thomas, G. E. (1999). Leaderless supervision and performance appraisal: A proposed research agenda. *Human Resource Development Quarterly, 10*, 91-94.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25-29.

Tinsley, H., & Weiss, D. (1975). Interrater agreement and agreement of subjective judgments. *Journal of Counseling Psychology, 22*, 358-376.

Townsend, A. M., & Scott, K. D. (2001). Team racial composition, member attitudes, and performance: A field study. *Industrial Relations, 40*, 317-337.

Tsui, A. S., & Barry, B. (1986). Interpersonal affect and rating errors. *Academy of Management Journal, 29*, 586-599.

Tsui, A. S., Egan, T. D., & Xin, K. R. (1995). Relational demography: The missing link in vertical dyad linkage. In M. N. Ruderman (Ed.), *Diversity in work teams: Research paradigms for a changing workplace* (pp. 97-129). Washington, DC: American Psychological Association.

Tsui, A. S., & Gutek, B. A. (1999). *Demographic differences in organizations: Current research and future directions*. New York, NY: Lexington Books/Macmillan, Inc.

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

Tsui, A. S., & O'Reilly, C. A. (1989). Beyond simple demographic effects: The importance of relational demography in superior-subordinate dyads. *Academy of Management Journal, 32*, 402-424.

Turner, J. C. (1987). *Rediscovering the social group: A self-categorization theory*. Oxford: Blackwell.

van Knippenberg, D., & Schippers, M. l. C. (2007). Work Group Diversity. *Annual Review of Psychology, 58*, 515-541.

Van Velsor, E., & Leslie, J. B. (2001). Selecting a multisource feedback instrument. In et al. (Ed.), *Handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (pp. 63-78). San Francisco, CA: Jossey-Bass.

Van Velsor, E., Taylor, S., & Leslie, J. B. (1993). An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness. *Human Resource Management, 32*, 249- 263.

Wagner, W. G., Pfeffer, J., & O'Reilly, C. A. (1984). Organizational demography and turnover in top-management groups. *Administrative Science Quarterly, 29*, 74-92.

Waldman, D. A., & Avolio, B. J. (1991). Race effects in performance evaluations: Controlling for ability, education, and experience. *Journal of Applied Psychology, 76*, 897-901.

Walker, A. G., & Smither, J. W. (1999). A five-year study of upward feedback: What managers do with their results matters. *Personnel Psychology, 52*, 393-423.

Wayne, S. J., & Liden, R. C. (1995). Effects of impression management on performance ratings: A longitudinal study. *Academy of Management Journal, 38*, 232-261.

Webb, N. M., Shavelson, R. J., Kim, K., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinist mates. *Military Psychology, 1*, 91-110.

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

Williams, H. M. (2004). Measuring gender composition in work groups: A comparison of existing methods. *Organizational Research Methods, 7*, 456-474.

Zedeck, S. (1995). Review of Benchmarks. In *the twelfth mental measurements yearbook* (pp. 128-129): Retrieved February 21, 2006, from the EBSCOhost Mental Measurements database.

Zenger, T. R., & Lawrence, B. S. (1989). Organizational demography: the differential effects of age and tenure distributions on technical communication. *Academy of Management Journal, 32*, 353-376.