8-2004

What Do Self and Peer Ratings Really Measure?

Gary J. GREGURAS
Singapore Management University, garygreguras@smu.edu.sg

Chet Robie
Wilfrid Laurier University

Robert J. Koenigs
SYMLOG Consulting Group

Marise Born
Erasmus University Rotterdam

What do Self and Peer Ratings Really Measure?

Gary J. Greguras

Singapore Management University

Chet Robie

Wilfrid Laurier University

Robert J. Koenigs

SYMLOG Consulting Group

Marise Born

Erasmus University Rotterdam

Gary J. Greguras, Singapore Management University, School of Business, BIZ #03-05, 469 Bukit Timah Road, Singapore 259756. Chet Robie, Wilfrid Laurier University, School of Business and Economics, Waterloo, Ontario, Canada N2L 3C5. Robert J. Koenigs, SYMLOG Consulting Group, 18580 Polvera Drive, San Diego, CA 92128. Marise Born, Erasmus University Rotterdam, Institute of Psychology, Department of Social Sciences, P.O. Box 1738 Room J505, NL-3000 DR Rotterdam, The Netherlands.

Abstract

Using Kenny's (1994) social relations model (SRM), data from 29 organizational teams who completed developmental performance ratings of themselves and other team members were analyzed. The current study (a) estimated the amount of variance associated with the ratee, rater, and relationship components, (b) projected dependability estimates under different measurement conditions, and (c) explored the social and relational processes associated with group dynamics and functioning. Results indicated that the relationship component accounted for significantly more variance than the ratee component, which in turn accounted for significantly more variance than the rater component. In general, results also indicated dyadic and generalized reciprocity for some performance dimensions. Further, self-ratings were related to how one rates, and is rated by, others.

What do Self and Peer Ratings Really Measure?

Rapid changes in how organizations are structured and function have forced human resource practitioners to adjust their practices to meet the demands of today's organizations. Current trends indicate that organizations are becoming flatter and more decentralized (Murphy & Cleveland, 1995) and that many organizations have shifted from individually-based work to team-based work (Fedor, Bettenhausen, & Davis, 1999; Reilly & McGourty, 1998). Associated with these changes have been changes in how employee performance is measured, developed, and evaluated. In particular, many organizations now use self and peer ratings as part of their performance management systems (Church & Bracken, 1997).

For team based-work, peer and self-ratings may be the only sources capable of providing relevant information regarding group processes and individual contributions (Fedor et al., 1999). Peers may be an especially valuable source of performance information because peers generally work closely together, interact frequently, and have the opportunity to observe both task and interpersonal behaviors (Murphy & Cleveland, 1995). This proximity and frequency of interaction is predicted to increase rating accuracy (e.g., Wherry & Bartlett, 1982). However, it is precisely this proximity and frequency of interaction that also raises concerns over the use of peer ratings. Those skeptical of using peer ratings suggest that such ratings will be contaminated with friendship biases (Doll & Longo, 1962; Landy & Farr, 1983) and that peers may be unwilling to differentiate among one another in an attempt to maintain harmony within the group (Murphy & Cleveland, 1995). Ironically, although these frequent and close interactions have been argued to both improve and hinder the quality of peer ratings, such interactions have been largely neglected in the extant research (Hennen, 1997).

Early research on performance ratings focused on rating scale format in an attempt to increase the psychometric properties of the scales. Landy and Farr (1980) called for a moratorium on rating scale research and instead encouraged researchers to consider the cognitive processes of the raters. In response, research quickly shifted to studying cognitive processes, but similar to the research on rating scale format, this cognitive stream of research has been criticized on the grounds of only making minor contributions (Murphy & Cleveland, 1995) and of not advancing organizational practices (Banks & Murphy, 1985). As Murphy and Cleveland (1995) note, what appears to be missing in these earlier approaches, and what may explain their limited utility, is a consideration of social and contextual factors that influence performance ratings. Specifically, previous models have neglected to recognize that appraisals take place in a dynamic context (Murphy & Cleveland, 1995) in which rater and ratee interactions and relationships are important considerations. As Vance, Winne, and Wright (1983) note: "Reproduceable level effects attributable to raters or ratees only partially explain ratings. A dyad-based analysis capable of representing rater-ratee interactions may be necessary to fully explicate the process" (p. 619).

As called for by Vance et al. (1983), the current study focuses not only on rater and ratee effects, but also on rater-ratee interactions and relationships. In doing so, the current study addresses three issues raised by Murphy and DeShon (2000a) in need of future research to more fully understand what performance ratings represent. Specifically, the first objective is to use Kenny's (1994) Social Relations Model (SRM) to investigate self and peer ratings. Unlike classical test theory which only decomposes variance into true and error variance, SRM analyses permit target (ratee), perceiver (rater), and relationship (rater by ratee interaction) variance components to be estimated separately. As noted by Murphy and DeShon (2000a), a major

limitation of past research with respect to variance partitioning has been its inability to disentangle the perceiver effect from the relationship effect. Previous research (e.g., Greguras, Robie, Schleicher, & Goff, 2003), for example, has observed that a substantial amount of variance could be attributed to *some* function of the rater but was unable to identify if this variance was primarily attributable to the rater main effect or rater by ratee interaction effect. Because the current study is able to separate these effects, it is able to more precisely identify and estimate the amount of variance attributable to each of these variance components.

The second primary objective of the current study is to use the estimated variance components to project dependability (reliability) estimates under various measurement conditions. As Murphy and DeShon (2000a) note: "…it would be interesting to know the effect of including or excluding the rater variance in the error term consistent with the distinction between absolute and relative error in generalizability theory. This information could be used to inform the researcher about the exchangeability of raters in the rating process or used to determine the number of raters needed to achieve a particular level of generalizability for a particular research question" (p. 893). Previous studies were largely unable to provide these estimates because of their inability to separate rater from relationship effects.

The third primary objective of the current study is to merge and extend the organizational science literature on self and peer ratings with the social psychological literature on interpersonal perception. As discussed below, the SRM analyses estimate variance components attributable to the rater, the ratee, and their dyadic relationships. These variance components can then be correlated with one another and with self-ratings to provide information about the social and cognitive processes involved in the rating process. As Murphy and DeShon (2000a) note, the nested designs of the vast majority of studies on performance ratings make it virtually impossible

to sort out rater effects from relationship effects, and therefore, it is largely unknown if, and how, these effects might be correlated with one another. The current study overcomes this limitation by using a round-robin design to estimate and correlate the different variance components. We see this as a major strength of the paper as such analyses will allow for a greater understanding of self and peer ratings and rating processes. An overview of the SRM is presented next, followed by a discussion of its usefulness within the context of studying self and peer performance ratings.

Social Relations Model

The SRM developed out of the person-perception literature and provides both a theoretical basis and a statistical tool to assess the interdependencies among ratings (Kenny, 1994; Warner, Kenny, & Soto, 1979). When individuals evaluate the performance of their peers, their ratings likely are dependent on one another's ratings. For example, in a team setting, Jim's evaluations of Eric likely are related to Eric's evaluations of Jim. As Hennen (1997) notes, previous research on peer evaluations has ignored this nonindependence of data. By ignoring these interdependencies, results from typical statistical methods may be distorted (for a discussion, see Kenny & Judd, 1986), and importantly, meaningful information about the interdependencies among peers is lost. It is this very interdependency that makes the study of teams interesting and that is the focus of the current study. After all, if team members were not interdependent, members would merely represent a collection of individuals (Marcus, 1998). One advantage of the SRM is that it does not require an assumption of independence, but rather, the nonindependence among peers is of interest and is analyzed (Kenny, 1994; Marcus & Kashy, 1995).

The SRM is a special case of generalizability theory (Cronbach, Glaser, Rajaratnam, & Nanda, 1972) and decomposes perceptual variance on a construct into target effects, perceiver effects, and relationship effects. Target effects are considered true score variance and reflect a target's tendency to elicit similar ratings from all raters. Perceiver effects are rater main effects and reflect a rater's tendency to rate all persons similarly (e.g., leniently). Relationship effects are factors unique to specific dyads. To estimate target, perceiver, and relationship effects using the SRM, multiple perceivers must rate multiple targets. This requirement generally is met by employing a round-robin data collection design in which each member of a group rates, and is rated by, every other member. The round-robin design is distinct from more typical generalizability theory (GT) models in which multiple raters evaluate a given target but the given target does not evaluate those raters (Cronbach et al., 1972). The variance partitioning is conducted on each group and the components are then averaged across groups.

After the target, perceiver, and relationship effects have been estimated, these effects may be correlated with each other to try to help explain social psychological perceptual phenomena. For example, consider two teammates, Jim and Eric, and two performance dimensions, dependability and cooperation. A positive target-target correlation indicates that the ratee is seen similarly across performance dimensions. Using the above performance dimensions for illustration, a positive target-target correlation would indicate that if Jim is seen as dependable he is also seen as cooperative. A positive perceiver-perceiver correlation indicates that raters saw the targets similarly across the two performance dimensions. For example, if Jim sees Eric as dependable, he also sees Eric as cooperative.

Relationship effects may also be correlated to investigate dyadic reciprocity. Recall that the relationship effect represents a peer's unique relationship with another peer. A positive

relationship (e.g., Jim's relationship effect with Eric) – relationship (e.g., Eric's relationship effect with Jim) correlation indicates, for example, that if Jim rates Eric as more dependable than he rates others and more than others see Eric as dependable, then Eric is likely to rate Jim as more dependable than he rates others and more than others see Jim as dependable. Stated differently, a positive relationship-relationship correlation indicates that Jim tends to rate Eric higher on dependability than would be expected by perceiver and target effects, and Eric also tends to rate Jim higher on dependability than would also be expected by perceiver and target effects. Whereas dyadic reciprocity investigates the correlation between relationship effects, generalized reciprocity is at the individual level of analysis and investigates the correlation between perceiver and target effects. Generalized reciprocity indicates that, for example, if Jim sees others as dependable, he is also seen as dependable by others. The study of both dyadic and generalized reciprocity is a fundamental area of inquiry in social psychological perception (Kenny, 1994). For example, in a team setting, a finding of generalized reciprocity for a given construct may suggest a two-way elicitation process wherein the behavior of the ratee is eliciting similar (or dissimilar in the case of a negative correlation) behaviors from the raters. Note that perceiver and target effects are at the individual level and therefore cannot be correlated with the relationship effect which is at the dyadic level (Marcus, 1998).

In addition to correlating target, perceiver, and relationship effects with themselves or one another, target and perceiver variance components may be correlated with a team member's self-ratings (which are not used in estimating the variance components). Self-other agreement is assessed by correlating self-ratings with their target effects (e.g., are people who see themselves as dependable seen as dependable by others?). Similarly, one may correlate self-ratings with one's perceiver effects (e.g., do those who report being cooperative also think others are

cooperative?). Significant correlations between one's self-rating and one's perceiver effect is termed assumed similarity within the SRM framework. Because the relationship effect is at the dyadic level of analysis, it cannot be correlated with self-ratings. It is important to note how these self-target correlations differ from the existing self-other rating agreement research. Existing research typically correlates one's self-ratings with an average of ratings from a particular rater source (e.g., peers). In this typical approach, a mean of observer ratings contains target, perceiver, relationship, and error variance. As such, this typical approach results in a less precise estimate of self-other agreement than are self-other agreement indices calculated using target effects from the SRM which are not conflated with perceiver, relationship, and error variance components (Marcus, 1998).

Studies have applied the SRM to such diverse areas as leadership perceptions (cf. Kenny & Zaccaro, 1983; Malloy & Janowski, 1992), personality ratings (cf. Malloy & Kenny, 1986; Paulhus & Reynolds, 1995), group psychotherapy (cf. Marcus & Holahan, 1994; Marcus & Kashy, 1995), interpersonal attraction (cf. Kenny, Bond, Mohr, & Horn, 1996; Park & Flink, 1989), and juror influence (cf. Marcus, Lyons, & Guyton, 2000). However, few studies have applied the SRM to better understand performance ratings. A primary reason for the lack of application of the SRM in the area of performance measurement is that in most natural settings raters are nested within ratees, rather than the typical round-robin design employed in SRM studies. That is, historically supervisors evaluated their subordinates, but the subordinates would not evaluate one another or their supervisors. With the increased use of teams in organizations, however, peer ratings often lend themselves to round-robin designs. When peer ratings conform to a round-robin design, the SRM is a powerful approach to study the inherent social and relational processes associated with group dynamics and functioning. Studies using either GT or

the SRM to partition performance ratings into variance components are discussed below and are used to formulate hypotheses in the current study.

GT and Performance Ratings

Recall that the SRM is an application of GT to data gathered from reciprocal designs (Kenny, 1994). Like SRM, GT is capable of simultaneously estimating multiple sources of variance (e.g., variance due to ratees, raters) within a single, multi-faceted experiment. As reviewed below, the previous GT studies of performance ratings have generally investigated variance components associated with the rater, the task, and these components' interactions with the ratee (Clauser, Clyman, & Swanson, 1999).

Both Kraiger and Teachout (1990) and Webb, Shavelson, Kim, and Chen (1989) investigated self, supervisor, and peer performance ratings of military personnel made for research purposes. In both studies and for each rater source, results indicated that the target effect and the undifferentiated residual term accounted for a significant amount of variance in performance ratings. Note that, in the Kraiger and Teachout (1990) study, only one rater per source was available, and as such, rater effects and relationship effects could not be estimated but rather, variance attributable to these sources is included in the undifferentiated error term. In the Webb et al. (1989) study, multiple raters were only available for peers. Because peers were nested within ratees, the effects attributable to the rater main effect and rater interaction effects could not be computed separately but this combined effect accounted for approximately 25% of the total variance in peer ratings.

Greguras and Robie (1998) extended the above studies by analyzing the generalizability of performance ratings made for developmental purposes by supervisors, peers, and subordinates. Within-source analyses revealed, across sources, that the largest amount of

variance was attributable to an undifferentiated error term, followed by a combined rater main effect and rater by ratee interaction effect, followed by the ratee effect. These results are consistent with those of the two studies above. Extending the Greguras and Robie (1998) study, Greguras et al. (2003) analyzed the generalizability of subordinate and peer performance ratings made for developmental (Time 1) and administrative purposes (Time 2). For both purposes and rater sources, the largest source of variance was associated with the combined rater main effect and ratee by rater interaction effect, followed by the undifferentiated error term, followed by target effects. Like previous studies in this area, in both Greguras and Robie (1998) and Greguras et al. (2003), raters were nested within ratees thereby not permitting the rater main effect (perceiver effect) and the rater by ratee interaction effect (relationship effect) to be estimated separately.

The few studies that have used GT to analyze performance ratings have produced generally consistent results. Specifically, results from GT analyses presented above indicate that a substantial amount of variance in performance ratings is accounted for by target effects, a combined perceiver and relationship effect (combined because of the nesting of designs), and an undifferentiated residual term. One major limitation of the above studies has been their inability to separate the perceiver effect from the relationship effect (Murphy & DeShon, 2000a; Greguras et al., 2003). In contrast, one major advantage of the SRM is its ability to separate the perceiver from the relationship effects. Relevant SRM studies of performance ratings are discussed next.

SRM and Performance Ratings

Sullivan and Reno (1999) had students form groups which met weekly to work on a learning exercise. Self and peer evaluations were collected at two different time periods during the semester. Students rated group members on their predictions about how each group member

performed independently on the quiz, how much one contributed to the group's performance, and how much one liked the other person. Across all three measures and both time periods, the target effect accounted for a significant amount of variance (with 1 of the 6 effects being only marginally significant), and, on average, accounted for 42% of the variance in peer ratings. Across both time periods, the perceiver effect was significant for the predicted performance and contribution factors but not for the likeability dimension. On average the perceiver effect accounted for 17% of the variance in ratings. Because multiple indicators were available for the contribution factor, the relationship effect could be estimated apart from the residual error component. The relationship effect accounted for an average of 32%, and the residual error accounted for 8%, of the variance for this dimension.

Boldry and Kashy (1999) had student members of a university's Corp of Cadets evaluate team members across three performance dimensions (i.e., Motivation, Leadership, and Character) for research purposes. Individuals belonged to units of approximately 80 members each. Only freshman and juniors were recruited from these units (40 members each) and they rated their in-group members (same class) and out-group members (different class). Across groups and performance dimensions, the approximate amount of variance in ratings explained by target effects was 17%, by perceiver effects was 14%, by relationship effects was 19%, and by the residual effect was 51%.

Also using a student sample, Greguras et al. (2001) had students work together throughout the semester to complete a class project. Following the completion of the project, participants rated each other and oneself on six performance dimensions (e.g., cooperation). Results indicated that the target effect was significant for 4 of the 6 performance dimensions and on average accounted for 26% of the variance in peer ratings. Perceiver effects were significant

for all of the performance dimensions and on average accounted for 30% of the variance in peer ratings. Because Greguras et al. used single-item indicators of the performance dimensions, they could not estimate the relationship effect apart from the residual effect. This undifferentiated relationship and error variance term on average accounted for 44% of the variance in ratings.

Taken together, the emerging pattern of results across these three studies which have applied the SRM to performance ratings suggest that the target, perceiver, and relationship effects account for a significant amount of variance in performance ratings. These results are consistent with findings from the social psychological interpersonal perception literature (e.g., Frey & Smith, 1993).

Present Investigation

The three primary objectives of the current study are highlighted below and hypotheses are presented where appropriate.

Variance Components. The first objective of the current study is to use the SRM to partition the variance of peer ratings into target, perceiver, and relationship components. Combining results from both the GT and SRM studies suggest that a significant amount of variance in performance ratings is attributable to the target, perceiver, and relationship variance components.

H1: Target effects account for a significant amount of variance in peer ratings.

H2: Perceiver effects account for a significant amount of variance in peer ratings.

H3: Relationship effects account for a significant amount of variance in peer ratings.

Both Sullivan and Reno's (1990) and Boldry and Kashy's (1999) SRM studies found that the perceiver effect was smaller than both the target and relationship effects. These findings are

consistent with Kenny, Mohr, and Levesque's (2001) review of seven social psychological SRM studies that observed perceiver effects to be relatively small compared to target or relationship effects. As such, the current study hypothesizes:

H4: Target effects account for significantly more variance in peer ratings than do perceiver effects.

H5: Relationship effects account for significantly more variance in peer ratings than do perceiver effects.

Results from the two SRM studies capable of partitioning the perceiver effect apart from the relationship effect have been inconsistent with respect to whether the target or relationship effect is largest. Sullivan and Reno (1999) observed a larger target than relationship effect (based on one performance dimension), whereas Boldry and Kashy (1999) observed the reverse (for each of the three performance dimensions). Note that the participants in Sullivan and Reno's study were students from a class and may not have had sufficient time to fully develop relationships within their groups. Specifically, team members were given a group quiz to work on during class. As these authors note, the students could have worked on the quizzes independently and merely conferred answers. In contrast, participants in Boldry and Kashy were members of a Corp of Cadets and actually lived together. We suspect for intact teams that have worked together long enough to develop relationships and unique ways of interacting with one another, that the relationship variance component will be larger than the target effect.

H6: Relationship effects account for significantly more variance in peer ratings than do target effects.

Decision Study. The second objective of the current study is to use the estimated variance components from the SRM analyses to project dependability estimates under different

measurement conditions. In GT (again recalling that SRM is a special case of GT), a distinction is made between absolute and relative decisions. This distinction is important because how error is conceptualized and operationalized differs depending upon the type of decisions one wishes to make. Absolute decisions are made when an individual's absolute level of performance is being considered. For example, when managers decide which employees need additional training, absolute decisions likely are being made. In contrast, relative decisions are made whenever the rank-order of individuals is considered. For example, deciding which employees to promote likely requires a relative decision, because your decision is relative or dependent on all individuals being considered. The current study projects dependability estimates for different types of decisions (absolute versus relative) and under different measurement conditions (e.g., differing numbers of raters). As noted previously, this information provides information regarding the exchangeability of raters in the rating process and can inform practitioners developing peer rating systems regarding the appropriate number of peer raters needed to achieve dependable ratings.

SRM Intercorrelations. The third objective of the current study is to correlate the different variance components with one another and with self-ratings. As discussed above, these intercorrelations provide information about the perceptual processes of the raters and will provide a greater understanding of self and peer ratings.

Method

Participants

Participants included 29 teams of 8 individuals each for a total of 232 team members. Ratings were collected over a three year period between 1998 and 2000. Demographic information of the participants was not collected by the consulting organization to help protect

participant confidentiality. Teams represented a variety of organizations and functions including law enforcement, healthcare, technology, and manufacturing industries. Based on the number of teams and members within each team, this design has an estimated statistical power to estimate each of the variance components above .95 for medium effect sizes (see Lashley & Kenny, 1998 for a description of the program and analyses used to calculate statistical power for analyses using the SRM).

Measure

The measure used in the current study is part of the SYMLOG (A System for the Multiple Level Observation of Groups) approach to investigating group processes. The SYMLOG instrument contains 26 items. The frequency with which an individual exhibits the behaviors and values represented by an item is evaluated from *Rarely* = 0, to *Sometimes* = 1, to *Often* = 2. As a first step in the analyses, we conducted a principal factor analysis to examine the factor structure of the 26 items. Six factors were extracted with eigenvalues greater than 1. Moreover, the scree plot showed a clear break at the sixth factor. Simple structure was evidenced such that each item loaded heavily on only one factor. Given the simple structure and adequate reliability for the scales[1], we retained the six factor structure for the SRM analyses. The six factors are described below. The extracted factors are conceptually similar to performance dimensions used in previous peer rating, GT, and SRM studies. For purposes of comparisons, we note examples of some of the similarities below.

Teamwork Orientation. Teamwork Orientation is defined as the degree to which an individual accomplishes tasks by working with others and actively engages in behaviors for the common good of the team. Seven items loaded on this performance dimension and had an estimated reliability of $\alpha = .82$. A sample item is: Active teamwork toward common goals,

organizational unity. This dimension is similar to the Work Team Orientation scale found on Benchmarks, a multirater performance instrument (Lombardo & McCauley, 1994).

Individualism. Individualism is defined as the degree to which an individual is self-interested, self-protective, and self-oriented. Seven items loaded on this performance dimension and had an estimated reliability of $\alpha = .74$. A sample item for this dimension is: Tough minded, self-oriented assertiveness. This dimension is similar to the Selfishness trait in Boldry and Kashy (1999).

Rule Compliance. Rule Compliance is defined as the degree to which an individual adheres to group and organizational rules and norms in order to accomplish tasks. Three items loaded on this performance dimension and had an estimated reliability of $\alpha = .65$. A sample item of Rule Compliance is: Active reinforcement of authority, rules, and regulations. This dimension is similar to the Respectfulness of Authority trait in Boldry and Kashy (1999).

Dedication. Dedication is defined as one's commitment to the organization and its goals. Four items combined to create this scale and had an estimated reliability of $\alpha = .70$. A sample item is: Dedication, faithfulness, loyalty to the organization. Boldry and Kashy's (1999) peer rating instrument also included a Dedication factor.

Affiliation. Affiliation is defined as the degree to which one engages in behaviors that are friendly and cooperative in order to develop and maintain relationships. Three items loaded on this factor and had an estimated reliability of $\alpha = .65$. A sample item is: Popularity and social success, being liked and admired.

Motivation. Motivation is defined as the degree to which one is active/passive and does/does not engage in work related activities. Two items loaded on this factor and had a reliability estimate of $\alpha = .44$. A sample item from the Motivation scale is: Admission of

failure, withdrawal of effort. Note that for this scale higher scores represent less motivation. Boldry and Kashy's (1999) peer rating instrument also included a Motivation factor.

Procedure

Teams were from a variety of organizations and completed the instrument described above as part of an employee and team development program. Team members rated themselves and each team member using the 26-item instrument. The rating system helps people systematically think through how they perceive the performance and values of others and themselves. Participants only received aggregate feedback across raters of their scores on the items. These aggregate results were provided to each participant individually in a facilitated coaching session designed to protect confidentiality. In some cases, feedback was also given to the group as a whole. Ratings were used to aid in team development and not as a basis for administrative decisions (e.g., promotions).

Results

Data Structure and Analyses

In order to separate the relationship variance from the random error variance, multiple indicators of each performance dimension were required. That is, relationship variance can be separated from error variance by making multiple observations (rating multiple items) within a single interaction and then splitting these observations into segments (Kenny et al., 2001). As such, we split the 26 items into 20 manifest indicators (the computer program we used, SOREMO, only allows for 20 manifest indicators). For Teamwork Orientation and Individualism (the two scales with the most items) we created four indicators (i.e., 3 indicators were based on the average of two items each, and the fourth indicator was based on one item) and for the rest of the factors each item served as an indicator. Peer ratings were then

decomposed into target, perceiver, relationship, and error variance components by using Kenny's (1998) FORTRAN program SOREMO, which performs social relations analyses on data collected using a round-robin design. A detailed description of the program and its associated formulae may be found in Kenny (1994). We used Kenny and La Voie's (1984) between groups t-test to test whether the SRM parameter estimates were significantly different from zero. We used dependent groups t-tests with $g - 1$ degrees of freedom for the test of differences between SRM variances (for more information, see Kenny, 1998).

Variance Partitioning

The relative stable construct variance partitioning for the six dimensions is shown in Table 1. The relative variances indicate the percentage of variance of each rating that is attributable to each of the sources in the rating model. Across the six performance dimensions, the target effect accounted for approximately 13%, the perceiver effect accounted for 8%, the relationship effect accounted for approximately 17%, and the undifferentiated residual effect accounted for approximately 63% of the variance in performance ratings. Significance tests of the variance components indicated significant target, perceiver, and relationship variance components for each of the six performance dimensions. These results support Hypotheses 1, 2, and 3 that stated that target, perceiver, and relationship effects would all account for significant variance in peer ratings, respectively.

Hypotheses 4-6 predicted differences between variance components. These hypotheses were tested two ways. First, we tested these hypotheses by collapsing across scales. Consistent with SRM analyses, we calculated variance components for each performance dimension ($k = 6$) for each group ($g = 29$). We then summed across performance dimensions to derive an average and standard deviation for each variance component for each group. Hypothesis 4 predicted that

the target effect would account for significantly more variance in ratings than did the perceiver effect; this hypothesis was supported [$t(28) = 2.697$, $p < .05$]. Hypothesis 5 predicted that the relationship effect would account for significantly more variance in ratings than did the perceiver effect; this hypothesis was supported [$t(28) = -5.702$, $p < .001$]. Hypothesis 6 predicted that the relationship effect would account for significantly more variance in ratings than did the target effect; this hypothesis was also supported [$t(28) = -1.987$, $p < .05$, one-tailed].

In addition to collapsing across performance dimensions to test Hypotheses 4-6, we conducted similar analyses for each dimension. As can be seen from Table 1, performance dimension differences were observed. Specifically, the target effect was larger than the perceiver effect for 4 of the 6 performance dimensions. The relationship effect was larger than the perceiver effect for 5 of the 6 dimensions. Finally, the relationship effect was larger than the target effect for 4 of the 6 dimensions. When considering each dimension separately, Hypotheses 4-6 received only partial support.

Decision Study

The second main objective of the current study was to use the variance estimates to project dependability estimates under different measurement conditions. Recall that error is conceptualized and operationalized differently for relative and absolute decisions. Specifically, for relative decisions, each component that interacts with the ratee (not the rater main effect) contributes to the error variance, whereas for absolute decisions, each component's main effect and its interactions contribute to the error variance. As indicated in Table 2, the estimated dependability coefficients for relative decisions are always greater than those for absolute decisions when holding all other factors (e.g., number of raters) constant given that we observed

that a significant amount of variance was attributable to the perceiver effect (i.e., rater main effect).

Note that the estimates in Table 2 are for differing numbers of raters, but because SRM does not partition variance attributable to the item effect, we were unable to project dependability estimates at differing numbers of items using data from the current study. However, because practitioners likely are interested in how both the numbers of raters and items impact the dependability of observations, we conducted additional analyses drawing upon past research. Specifically, both Greguras and Robie (1998) and Greguras et al. (2003) used GT to analyze the dependability of peer feedback ratings. Their design did not permit the rater effect to be disentangled from the rater by ratee interaction effect, but their designs and analyses allowed the item and item by target interaction effects to be estimated. Averaging across those two studies, their results indicated that the item effect accounted for approximately 3.2%, and the item by target effect accounted for 3.4%, of the variance in peer ratings. We used these estimates for the item and item by target interaction effects and decreased the residual term accordingly. Using these estimates from previous studies allowed us to conduct another D-study wherein the dependability of peer ratings could be estimated with differing numbers of items and raters. Results from these analyses are shown in Table 3. Each participant in the current study was evaluated by seven raters and the average number of items across performance dimensions was 4.3 items. Results from this D-study suggest that at 7 raters and 5 items the dependability of observations for relative decisions is .73 and for absolute decisions is .67. Note that we were unable to estimate a rater by item and a rater by item by ratee interaction effect as none of the designs in the existing studies could estimate these effects.

SRM Intercorrelations

The third objective of the present investigation was to use the SRM framework to correlate variance components with one another and self-ratings to better understand self and other-perception with respect to performance ratings.

Generalized Reciprocity. Generalized reciprocity assesses how one generally sees others (rater main effect) and is seen by others (ratee effect) and does not consider the unique dyadic perspective (relationship effect). Generalized reciprocity is measured by correlating individual-level target and perceiver effects. Column 1 of Table 4 indicates that there were significant levels of generalized reciprocity for 3 of the 6 performance dimensions (i.e., Teamwork Orientation, Individualism, and Affiliation). Thus, those who are seen as evidencing high levels on these performance dimensions also tended to see others as evidencing high levels on these constructs. Interestingly, for Rule Compliance, a significantly negative correlation was observed between individual-level target effects and perceiver effects. That is, those who were seen as complying with rules tended to see others as not complying.

Dyadic Reciprocity. Dyadic reciprocity captures a dyad's unique way of relating or interacting with one another. Dyadic reciprocity is investigated by correlating the relationship effects from a dyad for a particular performance dimension. Inspection of column 2 of Table 4 indicates that for only the Teamwork Orientation and Affiliation performance dimensions were the dyadic relationship components correlated. These significant correlations indicate, for example, that Eric and Jim each sees one another as being higher on Teamwork Orientation than they each see other team members, and more than other team members see each of them.

Target-target Correlations. Target-target correlations investigate whether an individual is perceived similarly across performance dimensions. Typical correlations of performance ratings (rather than variance components) across dimensions are confounded with rater, relationship, and

error components. As such, correlating target variance components for different dimensions provides better estimates of the "true" relation between performance dimensions (Marcus, 1998). Table 5 presents the target-target correlations from the current study below the diagonal. For comparison purposes, we also computed correlations between performance ratings (not variance components) based on the average of the target's seven raters (i.e., we took an average of the 7 raters' ratings and correlated these across dimensions for each participant). As indicated by these correlations (see upper diagonal in Table 5), the target-target correlations were always larger than the correlations between the average peer ratings. This might be expected because the average peer ratings contain rater, relationship, and residual error components which likely attenuates the observed correlations.

Self-other Agreement. Self-peer agreement may also be assessed by correlating self-ratings with target effects. Again, the advantage of these analyses over the typical self-other agreement rating research is that the target effect is not conflated with perceiver, relationship, or residual effects. For 5 of the 6 performance dimensions, the correlations between self-ratings and target effects were significant (see Table 6). The self-target correlation was not significant for the Teamwork Orientation performance dimension. That is, team members generally saw themselves as they were seen by others except how they saw themselves for the Teamwork Orientation dimension. Similar to our previous analyses, in addition to computing a self-target correlation, we also computed a self-average other correlation by correlating one's self rating with the average peer rating received on that particular performance dimension (see Table 6). As evidenced in the table, the self-target correlations again are always larger than the self-peer average other correlations.

Assumed Similarity. Self-perceiver correlations assess whether how a rater sees oneself is related to how the rater sees others. As indicated in Table 6, for 5 of the 6 performance dimensions, significant correlations were observed (the only exception being for the Dedication dimension). Hence, in general, individuals tend to see others as they see themselves. We also correlated one's self-rating with one's average rating of one's seven peers for each performance dimension (see Table 6). With these analyses, again the self-perceiver correlations were larger than the self-other average rating of others (with the exception of the Dedication dimension).

Discussion

Recent estimates suggest that nearly half of all US organizations use some type of team management in their organizations (Devine & Clayton, 1999). The shift to team based work and other organizational changes (e.g., flatter organizations) have influenced changes in how employee performance is measured and developed. One change has been the increased use of self and peer ratings as part of an organization's performance management system. The current study used the SRM to analyze self and peer ratings to more fully explore what self and peer ratings actually represent. The use of the SRM to analyze performance ratings is in line with current recommendations to apply GT (SRM being a special case of generalizability theory GT) to individual and organizational level phenomena (e.g., Gerhart, Wright, McMahan, & Snell, 2000; Murphy & DeShon, 2000a, 2000b). Findings from each of the study's three main objectives are discussed below.

Variance Components

The current study partitioned rating variance attributable to the ratee, rater, rater by ratee interaction, and residual effects. Results indicated that each of these effects accounted for a significant amount of variance in peer ratings with the most variance attributable to the residual

(63%), followed by the rate by ratee interaction (17%), followed by the ratee (13%), followed by the rater effect (8%). The amount of relative variance attributable to each of these effects was similar to estimates in previous studies. For example, results indicated that the amount of variance attributable to the ratee effect in the current study (13%) is similar to previous GT (e.g., 13% in Greguras et al., 2003; 15% in Greguras & Robie, 1998) and SRM (e.g., 17% in Boldry & Kashy, 1999) estimates of target effects for peer ratings. Similarly, the current study and several SRM studies (e.g., Boldry & Kashy, 1999) have observed that the residual effect accounts for over half of the variance in peer ratings. Large residual effects suggest that additional factors not modeled in the current study likely account for significant variance in peer ratings. The congruence in estimates across this and existing studies suggest that these estimates may be robust across samples, settings, and instruments.

Unlike previous GT studies in this area, the current study was able to disentangle the rater main effect from the rater by ratee interaction effect. Estimates from the current study suggest the relationship effect accounts for 2.13 times the amount of variance accounted for by the rater effect, and 1.31 times the amount of variance accounted for by the target effect. Consistent with past assumptions, observing significant rater and relationship effects indicate that these effects are not random sources of error, but rather should be modeled as systematic sources of rating variation (Murphy & DeShon, 2000a). The significant relationship component indicates that members of dyads are actually relating to one another in unique ways. Contrary to the "tentative conclusion" (p. 879) drawn by Murphy and DeShon (2000a) that rater effects are probably similar in magnitude to, or larger than, the residual component, our results suggest that combined rater main effect and rater by ratee interaction effect account for significantly less variance in performance ratings than does the residual component.

Another finding was that the relative amount of variance attributable to each component differed across performance dimensions. Observing dimension differences is consistent with existing performance rating research (e.g., Borman, 1979; Viswesvaran, Ones, & Schmidt, 1996), yet the reasons for such differences are infrequently empirically tested. Several potential reasons for these differences have been hypothesized including raters having different values, experiences, training, opportunities to observe behavior, or standards for different performance dimensions (Murphy & Cleveland, 1995). Recent research indicates that raters often have different goals and that these differences translate into different ratings (Murphy, Cleveland, Skattebo, & Kinney, 2004). It might be that for some performance dimensions there is greater variation in rater goals and this increased variation results in differences in variance component estimates. Future research could assess these and other potential reasons for performance dimension differences by expanding the rating model used in the current study to include such factors, and then estimating the amount of variance attributable to these components.

It is also important to note that our sample was comprised of in-tact organizational teams, whereas, the few previous SRM studies on performance ratings have used student samples. Because one focus of the paper was on how dyadic relationships influence the quality of observed ratings, it was especially important that the sample be comprised of peers who have a shared history of working together and the expectation that they will continue to work together. That is, it likely takes time for peers to develop unique ways of relating and interacting with one another (Kenny et al., 2001). Additionally, the previous SRM studies collected ratings for either research or administrative purposes. In contrast, the ratings analyzed in the present study were part of a team development program. Rater motivation and resultant ratings likely differ based on the purpose of the performance ratings (Murphy & Cleveland, 1995). As such, we believe the

results of the present investigation likely are more generalizable to organizational settings where peer ratings are collected as part of organizations' on-going performance development systems than are results from studies with student samples rating for research or administrative purposes.

D-study: Projected Dependability Estimates

The second objective of the current study was to use the variance component estimates from the SRM analyses to project dependability estimates under differing measurement conditions. SRM analyses only partition variance into ratee, rater, relationship, and residual effects. Because we were interested in how the dependability of observations would be affected by both changes in numbers of items and numbers, we used estimates of item variance and item by person variance from previous studies. Results from this D-study suggest that, across performance dimensions, the 8 raters and 4.3 items (on average) used in the current study would achieve the recommended level of dependability of .70 (Nunally, 1978) for relative decisions but would be a bit under .70 for absolute decisions. The contribution of this D-study is that it provides estimates that may help practitioners developing peer performance ratings systems decide upon the appropriate number of raters or items to employ to achieve dependable ratings.

Two notes are worth mentioning regarding the D-study estimates. First, given that the estimated amount of relative variance attributable to each factor in the current study is similar to that of previous studies (e.g., Greguras & Robie, 1998), our D-study estimates for absolute decisions are necessarily similar to those of previous studies. For example, the D-study estimates for absolute decisions for peer developmental ratings in the current study are practically identical to those of Greguras and Robie (1998). This is encouraging as the results across the existing studies in this area seem to converge even though these estimates were based on different instruments, samples, and settings. Second, although previous studies have

estimated the dependability of ratings for relative and absolute decisions (e.g., Greguras et al., 2003), again they were unable to estimate the rater main effect apart from the relationship effect. Given the current study's ability to partition the rater main effect apart from the relationship effect, it is not surprising that the dependability estimates for relative decisions differed noticeably from estimates in previous studies. As such, it could be argued that the dependability estimates for relative decisions in the current study are better estimates than those of previous studies. Because most measurement applications in the behavioral and social sciences (Murphy & DeShon, 2000a), and work groups in particular (Ilgen & Feldman, 1983), rely on relative rather than absolute decisions, these relative estimates may be especially informative.

SRM: Intercorrelations between Self-ratings and Variance Components

The intercorrelations between the variance components and self-ratings provide information regarding social and relational processes associated with group dynamics and functioning. The current study first explored if perceptions of individuals within a group were reciprocated. We explored reciprocity at both the dyadic and individual levels. The only significant levels of dyadic reciprocity were for the Teamwork Orientation and Affiliation performance dimensions. Interestingly, these are the two dimensions that primarily tap interpersonal relationships and ways of interacting with other group members. Because peers develop unique relationships that likely influence their evaluations of one another, significant levels of dyadic reciprocity or "shared chemistry" (Kenny et al., 2001, p. 136) might be expected for such interpersonal dimensions. The norm of reciprocity (e.g., Gouldner, 1960) and social exchange theory (e.g., Thibaut & Kelley, 1959) also predict that how one behaves toward another is reciprocated. For the other performance dimensions where we did not observe a significant amount dyadic reciprocity, it may be that the behaviors comprising these dimensions

are not as easily reciprocated. This interpretation would be consistent with Kenny et al.'s (2001) suggestion that dyadic reciprocity might be more likely for prosocial (teamwork) or affective behaviors than other types of behaviors.

Kenny et al. (2001) notes that because in many studies there is so little partner variance, generalized reciprocity is rarely examined, but that when partner variance is found, generalized reciprocity (i.e., one generally views others as one is viewed) is also usually observed. We only observed generalized reciprocity for the Teamwork Orientation, Individualism, and Affiliation performance dimensions. On these dimensions, peers saw others as they were seen. Again the norm of reciprocity (e.g., Gouldner, 1960) and social exchange theory (e.g., Thibaut & Kelley, 1959) would predict such reciprocity. Unexpectedly, a significant negative target-perceiver correlation was observed for the Rule Compliance performance dimension indicating that those who were seen as complying with rules tended to see others as not complying.

Self-ratings have received considerable amount of research attention, especially recently given their increased use in employee development programs (e.g., Atwater, Ostroff, Yammarino, & Fleenor, 1998; Atwater & Yammarino, 1992). The correlations between self-ratings and target effects give some insight into the "accuracy" of self ratings if one defines accuracy in self-assessment as the relationship between self-ratings and the "true" component of observer ratings. The average self-target correlation in the present study was .43, suggesting moderate levels of agreement and suggesting that there is a degree of correspondence between how one sees oneself and how one is seen by others. Often self-ratings are correlated with an average peer rating. In the current study, the average self-average peer rating across performance dimensions was .27. Previous meta-analyses have estimated the average self-peer rating to be .19 in one study (Conway & Huffcutt, 1997) and .36 in another (Harris &

Schaubroeck, 1988). It is interesting to note that the average self-peer rating across those two meta-analyses is.27, the same as the correlation observed in the current study. Our results indicate that the actual correspondence between self and peer ratings is probably higher than estimated in these previous meta-analyses based on the self-target correlations in the current study.

The average self-perceiver correlation was .40 indicating that how one sees oneself is significantly related to how one sees others, referred to as assumed similarity in SRM terminology. Ross, Greene, and House (1977) have proposed that there is a "false-consensus bias" in that people assume that others think, feel, and behave as they do; our results support this proposal. Further, research suggests that increased acquaintance leads to greater assumed similarity and therefore the self-perceiver correlation increases with increasing acquaintance (Park & Judge, 1989). Given that the participants in the current study were part of in-tact organizational teams, observing significant self-perceiver correlations might be expected.

Limitations

There are several limitations of the current study. First, only one instrument from one team development program was investigated. However, the degree of similarity in estimates from the current study with those of previous studies suggest that our results are quite generalizable. Second, the SRM analyses did not permit the item effect and the item by interaction effect to be estimated. As such, we used estimates from previous studies for these components when conducting the D-study at differing numbers of raters and items. We believe that these are the best estimates given the current state of this literature, however, whether these estimates are appropriate for this sample or instrument is unknown. Third, although performance dimension differences were observed with respect to the variance estimates, additional data (e.g.,

rater goals) that may have helped to explore these differences were not available. Fourth, demographic information of the participants was not collected by the consulting organization. As such, we were limited in our ability to describe our sample, but again, the similarity of estimates from this study with existing studies suggest that the results are generalizable across samples.

Future Research

Several avenues for future research appear fruitful. First, studies should be designed to uncover the underlying processes that govern the large relationship effect found in the present study. Similarly, future research should investigate how these relationships develop and potentially change over time. It is likely as individuals become more acquainted they develop more unique ways of interacting with one another. Second, this study should be replicated with different performance instruments (e.g., performance-related constructs tied to a job-analysis) and conditions (e.g., peer evaluations used for administrative uses). Third, contextual characteristics (e.g., group size, support for the appraisal system, demographic similarity) should be identified that predict perceiver, target, and relationship variance components to provide a better understanding of the factors that influence self and peer ratings. Fourth, investigation of how the perceiver, target, and relationship variance components may be used to predict various individual (e.g., perceptions of appraisal fairness), group (e.g., team performance), and organizational (e.g., profitability) outcomes also should be pursued and are easily incorporated into the SOREMO program. Fifth, research should attempt to understand the performance dimensions differences observed in this and other studies.

Footnotes

[1]We were not overly concerned with the low reliability of the sixth factor because of the capability of the SRM to analyze only stable construct variance.

References

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

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

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology, 38,* 335-345.

Boldry, J. G., & Kashy, D. A. (1999). Intergroup perception in naturally occurring groups of differential status: A social relations perspective. *Journal of Personality & Social Psychology, 77,* 1200-1212.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64*, 410-412.

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

Clauser, B. E., Clyman, S. G., & Swanson, D. B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Assessment, 36,* 29-45.

Conway, J., M., and Huffcutt, A. I. (1997). Psychometric properties of multi-source performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10,* 331-360.

Cronbach, L. J., Glaser, G. C., Rajaratnam, N., and Nanda, H. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Devine, D. J., & Clayton, L. D. (1999). Teams in organizations. *Small Group Research, 99*, 678-712.

Doll, R. E., & Longo, A. E. (1962). Improving the predictive effectiveness of peer ratings. *Personnel Psychology, 15,* 215-220.

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

Frey, K. P., & Smith, E. R. (1993). Beyond the actor's traits: Forming impressions of actors, targets, and relationships from social behaviors. *Journal of Personality & Social Psychology, 65,* 486-493.

Gerhart, B., Wright, P. M., McMahan, G. C., & Snell, S. A. (2000). Measurement error in research on the human resources and firm performance: How much error is there and how does it influence effect size estimates? *Personnel Psychology, 53*, 803-834.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review, 25*, 161-178.

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., & Born, M. P. (2001). Applying the social relations model to self and peer evaluations. *Journal of Management Development, 20,* 508-525.

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

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

Hennen, M. E. (1997). *Consensus and meta-accuracy in self-managing work groups: A social relations analysis*. Unpublished dissertation, University of Connecticut, Stoors, CT.

Ilgen D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 5, pp. 141-196). Greenwich, CT: JAI.

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

Kenny, D. A. (1998). *SOREMO Version V.2*. [computer program]. Storrs, CT: University of Connecticut.

Kenny, D. A., Bond, C. F., Jr., Mohr, C. D., & Horn, E. M. (1996). Do we know how much people like one another? *Journal of Personality and Social Psychology, 71,* 928-936.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99,* 422-431.

Kenny, D. A., & La Voie, L. J. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 18, pp. 142-182). San Diego, CA: Academic Press.

Kenny, D. A., Mohr, C. D., & Levesque, M. J. (2001). A social relations variance partitioning of dyadic behavior. *Psychological Bulletin, 127,* 128-141.

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

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

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

Landy, F. J., & Farr, J. (1983). *The measurement of work performance: Methods, theory, and applications*. New York: Academic Press.

Lashley, B. R., & Kenny, D. A. (1998). Power estimation in social relations analyses. *Psychological Methods, 3,* 328-338.

Lombardo, M., & McCauley, C. (1994). *Benchmarks: A manual and trainer's guide.* Greensboro, NC: Center for Creative Leadership.

Malloy, T. E., & Janowski, C. L. (1992). Perceptions and metaperceptions of leadership: Components, accuracy, and dispositional correlates. *Personality and Social Psychology Bulletin, 18,* 700-708.

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

Marcus, D. K. (1998). Studying group dynamics with the social relations model. *Group Dynamics: Theory, Research, and Practice, 2,* 230-240.

Marcus, D. K., & Holahan, W. (1994). Interpersonal perception in group therapy: A social relations analysis. *Journal of Consulting and Clinical Psychology, 62,* 776-782.

Marcus, D. K., & Kashy, D. A. (1995). The social relations model: A tool for group psychotherapy research. *Journal of Counseling Psychology, 42,* 383-389.

Marcus, D. K., Lyons, P. M., Jr., & Guyton, M. R. (2000). Studying perceptions of juror influence In Vivo: A social relations analysis. *Law and Human Behavior, 24,* 173-186.

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

Murphy, K. R., Cleveland, J. N., Skateebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology, 89*, 158-164.

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

Murphy, K. R., & DeShon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology, 53*, 913-924.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Park, B., & Flink, C. (1989). A social relations analysis of agreement in liking judgments. *Journal of Personality and Social Psychology, 56,* 506-518.

Park, B., & Judd, C. M. (1989). Agreement on initial impressions: Differences due to perceivers, trait dimensions, and target behaviors. *Journal of Personality and Social Psychology, 56,* 493-505.

Paulhus, D. L., & Reynolds, S. (1995). Enhancing target variance in personality impressions: Highlighting the person in person perception. *Journal of Personality and Social Psychology, 69,* 1233-1242.

Reilly, R. R., and McGourty, J. (1998). Performance appraisal in team settings. In J. W. Smither (Ed.), *Performance appraisal: State of the art in practice* (pp. 244-277). San Francisco: Jossey-Bass.

Ross, L., Greene, D., & House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13,* 279-301.

Sullivan, M. P., & Reno, R. R. (1999). Perceiving groups accurately. *Group Dynamics: Theory, Research, and Practice, 3,* 1-10.

Thibaut, J., & Kelley, H. (1959). *The social psychology of groups*. New York: Wiley.

Vance, R. J., Winne, P. S., & Wright, E. S. (1983). A longitudinal examination of rater and ratee effects in performance ratings. *Personnel Psychology, 36,* 609-620.

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

Warner, R. M., Kenny, D. A., & Soto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology, 37,* 1742-1757.

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

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

Table 1

Relative Stable Construct Variance Partitioning for the Six Constructs

Construct	Target	Perceiver	Relationship	Residual	Total Absolute Variance
Teamwork	.154	.096	.238	.513	.173
Individualism	$.160_{a}$	.072	$.152_{a}$	.616	.138
Rule Compliance	$.090_{a}$	$.077_{a}$	.176	.656	.129
Dedication	.128	.060	.185	.627	.146
Affiliation	$.178_{a}$	.064	$.149_{a}$	.609	.186
Motivation	$.056_{b}$	$.104_{ab}$	$.106_{a}$	.734	.106
Across Dimensions	.128	.079	.168	.626	.146

Note. All variance components are significantly different from zero at $p < .01$. Number of groups (g) = 29. Number of individuals per group (n) = 8. Values with the same subscript are not statistically different ($p > .05$, one tailed). Relative variances are reported for ease of interpretation but the significance tests of the variance components were performed on the absolute variance components.

Table 2

Decision Study Estimates by Construct and Across Constructs

	Teamwork		Individualism		Rule Compliance		Dedication		Affiliation		Motivation		Across Constructs	
Raters	$\hat{\rho}^2$	$\hat{\phi}^2$	$\hat{\rho}^2$	$\hat{\phi}^2$	$\hat{\rho}^2$	$\hat{\phi}^2$	$\hat{\rho}^2$	$\hat{\phi}^2$	$\hat{\rho}^2$	$\hat{\phi}^2$	$\hat{\rho}^2$	$\hat{\phi}^2$	$\hat{\rho}^2$	$\hat{\phi}^2$
1	.17	.15	.17	.16	.10	.09	.14	.13	.19	.18	.06	.06	.14	.13
2	.29	.27	.29	.28	.18	.17	.24	.23	.32	.30	.12	.11	.24	.23
3	.38	.35	.38	.36	.24	.23	.32	.31	.41	.39	.17	.15	.33	.31
4	.45	.42	.45	.43	.30	.28	.39	.37	.48	.46	.21	.19	.39	.37
5	.51	.48	.51	.49	.35	.33	.44	.42	.54	.52	.25	.23	.45	.42
6	.55	.52	.56	.53	.39	.37	.49	.47	.58	.57	.29	.26	.49	.47
7	.59	.56	.59	.57	.43	.41	.52	.51	.62	.60	.32	.29	.53	.51
8	.62	.59	.63	.60	.46	.44	.56	.54	.65	.63	.35	.32	.56	.54
9	.65	.62	.65	.63	.49	.47	.59	.57	.68	.66	.38	.35	.59	.57
10	.67	.65	.68	.66	.52	.50	.61	.59	.70	.68	.40	.37	.62	.59
15	.75	.73	.76	.74	.62	.60	.70	.69	.78	.76	.50	.47	.71	.69
20	.80	.78	.81	.79	.68	.66	.76	.75	.82	.81	.57	.54	.76	.75

Note. $\hat{\rho}^2$ = generalizability coefficient estimate for relative decisions. $\hat{\phi}^2$ = index of dependability estimate for absolute decisions.

Table 3

Decision Study Estimates Across Constructs with Item Estimates

No. of raters	No. of items	$\hat{\rho}^2$	$\hat{\phi}^2$
1	1	.14	.13
1	3	.26	.22
1	5	.31	.26
1	10	.36	.29
2	3	.40	.35
2	5	.47	.40
2	10	.53	.45
3	3	.50	.43
3	5	.56	.50
3	10	.62	.54
4	3	.56	.50
4	5	.63	.55
4	10	.68	.61
5	3	.61	.54
5	5	.67	.60
5	10	.73	.66
6	3	.65	.58
6	5	.71	.64
6	10	.76	.69
7	3	.67	.60
7	5	.73	.67
7	10	.78	.72
8	3	.70	.63
8	5	.75	.69
8	10	.80	.74
9	3	.72	.65
9	5	.77	.71
9	10	.82	.76
10	3	.73	.66
10	5	.79	.72
10	10	.83	.78

Note. $\hat{\rho}^2$ = generalizability coefficient estimate for relative decisions.

$\hat{\phi}^2$ = index of dependability estimate for absolute decisions.

Table 4

Reciprocity Correlations for the Six Constructs

Construct	Target-Perceiver	Relationship
Teamwork	.21*	.36**
Individualism	.22*	.04
Rule Compliance	-.24*	.12
Dedication	.05	.12
Affiliation	.21*	.24*
Motivation	.06	-.06

Note. *p < .05. ** $p < .01$. Number of groups (g) = 29. Number of individuals per group (n) = 8.

Lower scores on the Motivation performance dimension indicate more motivation.

Table 5

Target – Target and Other-Other Correlations

Construct	Construct 1	Construct 2	Construct 3	Construct 4	Construct 5	Construct 6
Teamwork	1.00	-.49**	.17	.56**	.47**	-.09
Individualism	-.70**	1.00	-.33**	-.53**	-.21	-.33**
Rule Compliance	.23	-.55**	1.00	.41**	-.11	-.14
Dedication	.74**	-.72**	.60**	1.00	.07	-.09
Affiliation	.52**	-.34**	-.22	.07	1.00	.15
Motivation	-.19	-.57**	-.22	-.42*	.45*	1.00

Note. Correlations above the diagonal were calculated by averaging the ratings across the seven raters per target and then correlating these averages between performance dimensions. Correlations below the diagonal are target-target variance component correlations. Both sets of correlations were computed for each group and a one-sample t-test was used to test if the average correlation across groups was different from zero (Kenny, 1994). * $p < .05$. ** $p < .01$. Lower scores on the Motivation performance dimension indicate more motivation.

Table 6

Correlations of Self Ratings with SRM and Observed Components

Self-report	Target	Peer Average	Perceiver	Other Average
Teamwork	.11	.04	.42*	.24*
Individualism	.53**	.33*	.49**	.35*
Rule Compliance	.51**	.23*	.37*	.19
Dedication	.31*	.14	.17	.19
Affiliation	.68**	.39**	.29*	.17
Motivation	.43*	.23*	.64**	.31*

Note. *p < .05. ** $p < .01$. Number of groups (g) = 29. Number of individuals per group (n) = 8. Self-peer average rating correlations are correlations of self-ratings with the average ratings across the other seven peers of one's performance on that dimension. Self-other average rating correlations are correlations of self-ratings with one's average rating given to one's seven peers on that performance dimension. Lower scores on the Motivation performance dimension indicate more motivation.