

Factors Affecting the Convergence of Self–Peer Ratings on Contextual and Task Performance

Jennifer L. Mersman and Stewart I. Donaldson

*School of Behavioral and Organizational Sciences
Claremont Graduate University*

This study examines factors that predict the extent to which 408 operating-level workers rated themselves higher, lower, or the same as their coworkers rated them, for both task and contextual performance. On ratings of contextual performance, underestimators tended to be distinguished by significantly higher levels of both self-monitoring and social desirability. This trend operated similarly, though not significantly for task performance. Additionally, ratings of quantity of work obtained the highest degree of self–peer rating convergence as compared to ratings of quality of work and contextual performance. These results are discussed in terms of the practical implications for multirater systems.

Understanding the results of 360° feedback in performance appraisal is essential given the prevalence of these instruments. One issue that is particularly important is the congruence between self-ratings and other ratings (e.g., ratings from sources such as supervisors, peers, subordinates, and customers) because it affects how results from 360° feedback are interpreted and presented to participants. Considerable time and money is spent doing multirater feedback, and the convergence among raters is seen as a key variable in such systems (Yammarino & Atwater, 1997). The extent of rater congruence in multirater systems is of practical importance because it affects how results are interpreted and presented to participants. Users of such systems and practitioners that implement them spend much time in interpreting what the rating discrepancies mean for the ratee, and how the ratee should deal with this information. As Brutus, Fleenor, and Tisak (1996) noted, the

identification of discrepancies is an important part of the developmental process for the rater. The identification and interpretation of these rating discrepancies also has practical implications from a change management perspective. For example, it has the potential to help human resource practitioners identify individuals most likely to need extra training when a 360° feedback system is implemented. However, what has not been fully examined are the predictors of rating agreement, and when rating agreement should and should not be expected.

A greater understanding of the factors that influence the convergence of ratings will lead to a greater understanding of the construct of rater agreement. The purpose of this study is to investigate these potential individual difference variables as they influence self–other (S–O) agreement on ratings of contextual performance—a performance measure that is not typically employed in studies of rater agreement. However, before the importance of contextual performance and the issue of rating convergence on it can be discussed, it is first necessary to examine the importance of rating convergence itself.

VALUE AND MEANING OF CONVERGENCE

Given that correlations between self- and peer ratings range from .05 to .69 (Harris & Schaubroeck, 1988; Mabe & West, 1982), perhaps the most important question for understanding S–O rating agreement is what convergence actually means. Simply defined, convergence is the extent to which ratings from multiple sources are similar as determined by a direct comparison among them. This definition is easy to understand, but the underlying meaning of convergence has received much debate. For example, convergence between self- and other ratings of performance may be an indicator for convergent validity (Vance, MacCallum, Coovert, & Hedge, 1988), leniency bias (Williams & Levy, 1992), self-awareness (Atwater & Yammarino, 1992; Church, 1997; Van Velsor, Taylor, & Leslie, 1993; Wohlers & London, 1989), or accuracy (Yammarino & Atwater, 1993, 1997).

Yammarino and Atwater (1993, 1997) defined accurate ratings as ratings that are in agreement, and accurate estimators as those who rate themselves in alignment with how others rate them. This line of thought is consistent with what Bozeman (1997) labeled the “traditional” view of interrater agreement—that convergence leads to reliability, which subsequently leads to validity. Thus, it is typically believed that lack of agreement indicates invalid ratings.

In contrast to this traditional view Bozeman (1997) and others (e.g., Borman, 1974; Murphy & Cleveland, 1995) argued that interrater agreement may be a “nonissue” because different raters may be rating different aspects of performance and/or using different information in their evaluations. Different rating sources offer different perspectives, and this is where the utility of multisource rating lies. This being the case, we should neither automatically expect nor de-

sire convergence. In fact, a very high level of convergence could indicate that additional ratings offer redundant information and that collecting them is a waste of organizational resources. Alternatively, convergence could represent the correlation of bias—positive or negative.

Therefore, convergence is clearly not an indicator of “true score” or accuracy in all circumstances. Although interrater agreement among “others” (e.g., between peers, supervisors, or subordinates) may be neither desired nor expected, disagreement between self- and others raises some interesting questions about the reasons behind the discrepancy. This is where the issue of convergence and its implications is most relevant. Researchers should be concerned about discrepancies of S–O ratings because of what these discrepancies may mean in terms of individual differences (what types of ratees are likely or not likely to rate themselves in alignment with how others rate them); practitioners should be concerned because important developmental and performance outcomes are dependent on the interpretation of these discrepancies.

DIRECTION OF DISAGREEMENT

Disagreement between self-ratings and other ratings can be in either two directions: self-ratings greater than other ratings (i.e., ratee is an overestimator), and self-ratings less than other ratings (i.e., ratee is an underestimator). This direction of rater agreement may be suggestive of important individual differences. Yammarino and Atwater (1997) presented a framework of consequences for human resource management depending on the extent to which S–O ratings converge. They posit that more negative outcomes result when self-ratings are significantly higher than the ratings of others. For example, overestimators may tend to misdiagnose their strengths and weaknesses and fail to see the need for training and development. Therefore, failure to understand the reasons behind rating discrepancies may result in misplaced coaching or missed opportunities for performance improvement.

However, it is not necessarily the case that overestimators will remain overestimators. This is important in light of a finding by Atwater, Roush, and Fischthal (1995) that suggests not only can feedback be useful for overestimators, but that learning to accurately rate oneself may be predictive of performance. Atwater et al. found that when overestimators received feedback that their self-ratings were higher than the ratings of others, overestimators tended to adjust their performance accordingly. In this study, 32 weeks after the initial rating and feedback occasion, leaders were rated again (by subordinates and themselves). Overestimators gave lower self-ratings than the first rating occasion and also received higher subordinate ratings, thus indicating a positive change in performance. Although self-ratings for underestimators increased, subordinates’ ratings

on the second rating occasion did not differ from the first. As this study suggests, the direction of disagreement may be important for behavioral and performance outcomes; therefore it is necessary to understand the factors that predict rater agreement as well as the direction of rater disagreement (in the form of over and underestimation), and whether these factors differ.¹

CORRELATES OF CONVERGENCE

A number of researchers have looked into this issue and have made some interesting observations regarding the implications of rater agreement. Several studies have found congruence between self- and other ratings to be correlated with performance (Atwater & Yammarino, 1992; Bass & Yammarino, 1991; Van Velsor et al., 1993). More recently, Furnham and Stringfield (1994) found that managers rated by their superiors as more successful tended to have fewer discrepancies between their self- and subordinates' ratings. Similarly, Church (1997) found that the degree of agreement between S-O ratings significantly distinguished average performers from high performers as measured by performance history, achievements, and supervisor evaluations. Put simply, the evidence suggests that S-O ratings are more similar for high performers.

Given that rating convergence is related to performance, it is important to identify the factors that potentially affect the extent of agreement between S-O ratings, as well as the direction of the disagreement. Researchers have begun to describe these factors (see Table 1). For instance, there is evidence to suggest that job type (Harris & Schaubroeck, 1988), job level (Brutus, Fleenor, & McCauley, 1996), type of organization (Wohlers, Hall, & London, 1993), and aspects of the performance appraisal system itself (Schrader & Steiner, 1996) tend to affect the degree of S-O rating convergence. There is also research to suggest that individual characteristics, such as knowledge of the appraisal system (Williams & Levy, 1992), Dominance (Brutus et al., 1996), Introversion (Fleenor, Brutus, & McCauley, 1996), and self-monitoring (Church, 1997), influence the amount of S-O rating agreement obtained in multirater systems.

Although this research is a good beginning, ratee characteristics that impact rater agreement have not been fully examined. To better understand what rating convergence is and what it implicates, it is necessary to continue to explore additional ratee individual difference and personality variables as they predict convergence. Moreover, it is important to examine convergence across multiple measures of job performance that include the task and contextual performance domains.

¹As one anonymous reviewer pointed out, the factors that predict agreement may not be the same as the factors that predict disagreement.

TABLE 1
Summary of Research on Correlates of Convergence

<i>Level of Analysis</i>	<i>Research Findings</i>
Individual	<p>Performance: Greater convergence associated with higher performance (Atwater & Yammarino, 1992; Bass & Yammarino, 1991; Church, 1997; Furnham & Stringfield, 1994; Van Velsor, Taylor, & Leslie, 1993).</p> <p>Knowledge: Greater convergence associated with greater rater knowledge of appraisal system (Williams & Levy, 1992).</p> <p>Personality: Low Dominance characterizes underestimators, and high Dominance characterizes overestimators (Brutus, Fleenor, & McCauley, 1996); low Introversion differentiates overestimators from underestimators (Fleenor, Brutus, & McCauley, 1996); high self-monitors obtained greater degree of convergence (Church, 1997).</p>
Job	<p>Job level: Higher managerial levels obtain less convergence (tend to overestimate performance relative to others' ratings) than lower level managers (Brutus, Fleenor, & Tisak, 1996).</p> <p>Job type: Greater convergence obtained in blue collar/service jobs than for managerial or professional jobs (Harris & Schaubroeck, 1988).</p>
Organizational	<p>Organizational Type: Convergence more likely to be obtained in private firms than government organizations (Wohlers, Hall, & London, 1993).</p> <p>Features of Appraisal System: Greater convergence associated with more explicit and objective comparison standards for rating (Schrader & Steiner, 1996).</p>

CONVERGENCE AND CONTEXTUAL PERFORMANCE

As opposed to task performance (performance defined by job descriptions and formally rewarded) contextual performance includes behaviors that are neither outlined for nor expected of an employee (Borman & Motowidlo, 1993). Behaviors in contextual performance are labeled *citizenship behaviors* by Organ (1988), who defined organizational citizenship behavior (OCB) as individual behavior that in the aggregate aids organizational effectiveness.² This behavior is neither required by the individual's job description, nor directly rewarded by a formal reward system, and as such can be thought of as extra-role performance. Smith, Organ, and Near (1983) demonstrated that there are two factors in the OCB scale: altruism and generalized compliance. *Altruism* describes the employee who directly and intentionally helps individuals in personal interactions. *Generalized*

²Organ (1997) noted that the terms organizational citizenship behavior (OCB) and contextual performance are essentially synonymous, but that the term OCB may carry more implicit meaning. In this article, the term contextual performance will be used when referring to the overarching construct of OCB/contextual performance, but citizenship behaviors or OCB will be used in referring to the measurement of that construct. This treatment is justified given that OCB is viewed as one component within the larger construct of contextual performance (Borman & Motowidlo, 1997).

compliance refers to an impersonal form of conscientiousness manifested by adherence to internalized norms.

The notion of contextual performance is important to fully describe the criterion domain of job performance (Borman & Motowidlo, 1993). However, studies in rater agreement typically use task performance as the measure on which rating convergence is assessed. The extent of rater agreement using measures of contextual performance has not been adequately addressed in research to date. Moreover, the question of whether or not convergence differs according to performance dimension—task or contextual—has not been fully explored (cf. Conway, 1996). Although research has examined the stability of rating agreement across different aspects of performance, like leadership skills, communication, and judgment (Nilsen & Campbell, 1993), research on rating agreement across less verifiable measures of performance such as OCB is lacking.

As the criterion space for job performance is expanded to include citizenship behaviors, it is necessary to understand the S–O rating processes associated with this measure of performance, especially as they relate to or differ from the rating processes on more traditional performance measures. Therefore, in an effort to investigate the factors that account for variance in rating agreement on contextual performance, this study examines two individual difference variables as they predict convergence of ratings on contextual performance: self-monitoring (SM) and social desirability (SD).

SM

Snyder (1974) described SM as a construct that refers to the control of self-presentational behaviors. People high in SM attend to situational cues that guide their self-presentation of what they believe to be appropriate behaviors. In contrast, low self-monitors only display behaviors consistent with their true feelings, regardless of situational cues. S–O rating agreement is related to SM. Church (1997) showed that managers higher in SM tended to show more agreement between self-ratings and ratings from direct reports. This attention to and regulation of one's own behavior may lead the high SM ratee to be better able to accurately assess his or her own behaviors, and consequently have self-ratings that are more in alignment with the ratings of others.

Because high self-monitors tend to change their behavior to fit the situation, they are more aware of the behaviors they present in each social context. This tendency for high self-monitors to observe and control their own behavior may make past behaviors more easily accessible and available for recall in a performance-rating context. Indeed, higher levels of SM would tend to be less associated with what Ashford (1989) discussed as problems in the task of self-assessment. One difficulty in the task of self-assessment is inattention to rel-

evant cues in the situation: "The difficulty individuals face, then, is that they must judge which verbal messages, situations, behaviors, or lack of behaviors are relevant and should be considered feedback cues" (Ashford, 1989, p. 146). High self-monitors are more adept in reading these situationally dependent cues, and should therefore have a better established behavior-feedback cue link than low self-monitors.

Church (1997) discussed SM as a counterpart to the construct of self-awareness. The ability to monitor and adjust one's own behavior directly corresponds to one's level of self-awareness. This congruence between SM and self-awareness is also found in Wicklund's (1975) conceptualization of self-awareness. He defined self-awareness as self-focused attention, or as an individual's ability to self-observe. Given Church's conceptualization of SM and his findings, as well as the logic outlined in this study, it is reasonable to suggest that SM will lead to more congruent self-peer ratings via the self-awareness connection.

Hypothesis 1a: There will be significant differences in SM among individuals in the various self-peer rating congruence categories based on ratings of contextual performance. Specifically, higher levels of SM will characterize in-agreement raters.

Hypothesis 1b: In-agreement raters will likewise be characterized by higher levels of SM when the rating categories are based on task performance.

SD

SD refers to the tendency to respond in such a manner so as to make oneself look good, and is taken to reflect a socially conventional and dependable persona (Paulhus, 1991). Although some SD instruments measure impression management (e.g., Paulhus', 1984, Balanced Inventory of Desirable Responding), the Marlowe-Crowne Social Desirability Scale (MCSD; Crowne & Marlowe, 1960) has been interpreted as assessing the need for approval (Crowne & Marlow, 1964; Paulhus, 1991). Thus, although some measures of SD tend to assess the tendency to positively dissimulate, the MCSD tends to assess one's need to be considered honest or likable in order to receive social approval (Carver & Scheier, 1992; Paulhus, 1991).

Defining the SD construct as need for approval is fitting in a performance evaluation context where it may be socially desirable to be modest about self-ratings. Otherwise the self-ratings may be perceived as overly positive, and consequently the self-rater could receive social disapproval due to rating dishonesty. Those with a high need for approval (i.e., receiving higher scores on the MCSD) would have more motivation to give accurate self-assessments in the situation where they know that their ratings will be compared to those of their peers. In this study, this is

TABLE 2
Summary of Agreement Categories Based on Ratings of Contextual Performance

	<i>Underestimators</i>	<i>In-Agreement Raters</i>	<i>Overestimators</i>
Self-monitoring	Low	High	Low
Social Desirability	High	High	Low

TABLE 3
Summary of Agreement Categories Based on Ratings of Task Performance

	<i>Underestimators</i>	<i>In-Agreement Raters</i>	<i>Overestimators</i>
Self-monitoring	Low	High	Low
Social Desirability	High	High	Low

true of half the sample because a counter-balanced design was employed (half the participants completed ratings on their peers before completing ratings on themselves, and therefore knew when rating themselves that their peers would likewise be rating them, see Methods section).

Hypothesis 2a: There will be significant differences in SD among individuals in the various self-peer rating congruence categories based on ratings of contextual performance. Specifically, in-agreement raters and underestimators will be characterized by higher levels of SD than overestimators.

Hypothesis 2b: In-agreement raters and underestimators will likewise be characterized by higher levels of SD when the rating categories are based on task performance.

The relation of SM and SD with rating convergence is examined across both contextual and task performance because it is important to determine whether these relations operate similarly or differently depending on the measure of performance being rated (see Table 2 and Table 3 for a summary of Hypotheses 1 and 2). Examining the convergence of ratings on contextual performance is necessary because although citizenship measures are becoming more relevant given the changing nature of work, much research to date has failed to explore the rating convergence process using this measure of performance.

CONTEXTUAL PERFORMANCE AND VERIFIABILITY

In addition to rater and ratee characteristics, the type of performance measure used for rating may have an effect on congruence. As previously noted, studies

in rating agreement typically use measures of task performance. It is reasonable to suggest that the degree of rating congruence may vary when using measures of task performance as opposed to measures of contextual performance. With task performance it may be more likely that both rater and ratee have access to the same, unambiguous performance information. As such, greater rater agreement would occur. Examples of such verifiable performance measures are documented absenteeism and quantity of work performed. Less verifiable performance measures would include contextual performance (citizenship behaviors) and quality of work performed because these measures are subject to more individual interpretation. This is supported by Furnham and Stringfield (1998) who found more rating congruence when ratings were based on observable behaviors (e.g., task performance) compared to when they were based on less observable behaviors (e.g., cognitive variables).

Hypothesis 3: Greater convergence is predicted for more verifiable measures of performance such as quantity of work performed and attendance, whereas less agreement is predicted for quality of performance and overall task performance, and even less for contextual performance.

METHOD

Participants

The data used for this study are taken from Project WORKWELL conducted in California (Donaldson, Ensher, & Grant-Vallone, 2000). This project was designed to investigate variables of employee lifestyle and mental health as they predict important organizational outcomes (e.g., job satisfaction, performance, OCB, absenteeism, etc.). Moreover, information was collected on each participant from a coworker who knew the participant.

There were 408 participants in the study (a result of 204 coworker pairs). Of these, 68% were women, 25% Latino, 22% European American, 16% African American, and 10% Asian. Employees were all operating level employees who worked an average of 38 hr per week ($SD = 11.2$; $Mdn = 40$). The average length of tenure with their companies was 4.8 years ($SD = 5.25$; $Mdn = 3$). Participants worked in a broad range of occupations, and came from a wide variety of organizations. More than half of the participants reported being single (56%), whereas 26% reported being married, 13% divorced, and 2% widowed. Of the participants, 50% reported their annual personal income as being less than \$20,000, whereas 87% reported a total annual personal income lower than \$30,000. More than 87% of the employees reported an education level of less than a 2-year, junior college degree.

Measures

Contextual Performance. OCB was assessed using the Smith et al. (1983) measure. For the self-measure, the items were worded in the first person, and on the peer report the items were worded to indicate, "My coworker . . ." This scale is broken down into two factors: Altruism and Generalized Compliance. Smith et al. report a coefficient alpha of .88 for Altruism and .85 for Generalized Compliance.

For the purposes of this study, only the Altruism scale was used to assess convergence. This decision was made due to the results of a factor analysis of the OCB measure. A principal components analysis was conducted suggesting a two-factor solution. This two-factor solution was then forced using the principal axis factoring method with oblique rotation. Based on an examination of the factor loadings in the structure matrix, three items (6, 9, and 11) were removed from further analysis because they cross-loaded on more than one factor. The remaining 13 items were then analyzed using the principal axis factoring method with oblique rotation. This resulted in a cleaner more interpretable solution with no cross loadings greater than .3. Factor one is labeled *Altruism* and consists of Items 1, 3, 5, 7, 12, 13, and 15. Factor two is *Generalized Compliance* and is composed of Items 2, 4, 6, 10, 14, and 16. The interfactor correlation matrix showed that these two factors are somewhat correlated ($r = .318$). However, an analysis of these items reveals that Generalized Compliance is probably more indicative of the manner in which one spends one's time at work according to what is expected. This may be interpreted more along the lines of task behavior rather than contextual performance. For instance, scoring low on this factor would indicate that an individual does spend time in idle conversation, on the phone, and taking undeserved breaks. Across organizations in general, these behaviors can result in some sort of sanction if not at the very least the withholding of formal rewards; citizenship behavior is defined as behavior that is not required, expected, or rewarded (Organ, 1988).

A coefficient alpha of .77 was obtained for self-ratings of Altruism, whereas a coefficient alpha of .79 was obtained for peer ratings. These reliabilities are less than those reported by Smith et al. (1983).

SM. SM was assessed by Snyder's (1974) Self-Monitoring scale. In this sample, items for this scale were scored on a Likert-type scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*) rather than a dichotomous true false format to offer more variance in the scores. This sample obtained a coefficient alpha of .75 for the overall factor of SM.

However, research suggests that the SM scale is multidimensional (Briggs, Cheek, & Buss, 1980; Gabrenya, & Arkin, 1980). Therefore a factor analysis was conducted on the scale and the subscale of Other Directedness (OD) was used in the analyses. After a principal components analysis suggested a minimum of a

three-factor solution, a principal axis analysis was conducted. The structure matrix revealed that 6 of the 25 items (1, 4, 6, 18, 20, and 25) had cross loadings greater than .3. The factors resembled the three factors discussed in Briggs et al. (OD, Extraversion, and Acting). After the remaining 19 items were analyzed again using principle axis, the factors were more interpretable with no cross loadings greater than .3. The first factor (Items 2, 3, 7, 15, 17, and 19) resembled OD and obtained a coefficient alpha of .60.

The decision to use OD and not the other subscales of Extraversion and Acting was based on theoretical and empirical considerations. First, of the three subscales OD most captures what is meant by the SM construct: sensitivity to social cues, self-regulation, and concern for the appropriateness of social behavior (Briggs et al., 1980; Lennox & Wolfe, 1984). Second, this factor aligns well with the logic of Hypotheses 1a and 1b—that people who have a tendency to monitor and control their own behavior may be better able to recall their performance because they pay particular attention to the expression of their behavior. Extraversion (being comfortable in social situations, being the center of attention, and telling jokes) and Acting (being good and acting, entertaining, and charades) do not adequately describe one who is sensitive to social cues and manages one's behavior based on those cues. And finally the OD scale was the cleaner, most interpretable factor with no cross loadings.

SD. SD was assessed using the MCSD (Crowne & Marlow, 1960). Reported alpha coefficients for this scale range from .73 to .88, whereas test-retest reliabilities range from .84 to .88 (Paulhus, 1991). The coefficient alpha obtained in this sample is .83.

Task Performance. Coworkers assessed the target's job performance according to five categories using a 5-point rating scale as follows: quantity of work, 1 (*very slow worker*) to 5 (*Rapid worker. Unusually big producer*); quality of work, 1 (*Poor Quality. Too many errors*) to 5 (*Exceptionally accurate, practically no mistakes*); overall performance, 1 (*Poor—Bottom 10%*) to 5 (*Superior—Upper 10%*); attendance, 1 (*Poor, absent too often*) to 5 (*Exceptional, perfect attendance*); promptness, 1 (*Poor. Usually late for work*) to 5 (*Exceptionally prompt. Always on time for work*). The task performance scale was created from the mean of these five items; its coefficient alpha for peer ratings was .76, and for self-ratings was .67.

Degree of Convergence. There are several methods for measuring convergence (for a complete review of these methods, see Brutus, Fleenor, & Tisak, 1996). Although a difference score approach seems most intuitive (taking the dif-

ference between self- and other ratings), it is not the best method statistically because of problems with the reliability of difference scores (i.e., the more reliable the components that make up the difference score, the less reliable that difference score; and the correlation between an unreliable measure and a criterion of interest will be attenuated). Further, using an interrater reliability approach is problematic given the instability of the correlations. Although these correlation coefficients between self- and other ratings can be transformed and used as a dependent variable (i.e., convergence), the reliability may be low if the number of items on which the correlations are based is small, as it is in this study (e.g., seven items for Altruism and five items for task performance). Caution must be used when interpreting correlations based on unreliable correlation coefficients as the criterion (Mersman & Shultz, 1998). Therefore, convergence was assessed by a three-category classification scheme simplified from a four-category scheme developed by Atwater and Yammarino (1992). Although some variance is lost by categorizing a continuous variable (Brutus et al.), this method is superior to using difference scores as the predictors or criteria. These three categories were developed by first standardizing the self- and peer ratings on Altruism. Individuals were then classified into one of three agreement categories: overestimators, underestimators, and in-agreement. Overestimators are those with a standardized self-Altruism rating (Z_{SAIt}) that is greater than one-half a standard deviation above the standardized peer Altruism ratings (Z_{PAIt}), underestimators are those with a Z_{SAIt} less the one-half a standard deviation of the Z_{PAIt} , and those classified as in-agreement have a Z_{SAIt} that lies between one-half a standard deviation above or below the Z_{PAIt} . This process for creating agreement categories was completed again using standardized ratings on the task performance scale. Thus, individuals were categorized as over-, under-, or in-agreement raters twice: once on Altruism and again on task performance. For Altruism, the group sizes for over-, under-, and in-agreement raters were 133, 151, and 110, respectively; for task performance, the group sizes for over-, under-, and in-agreement raters were 100, 108, and 133, respectively.

Procedure

Individuals who offered referrals for potential participants were offered \$5 for each coworker pair they recruited. The recruitment materials advertised \$50 total compensation for participation (\$25 at Wave 1, and \$25 at Wave 2), a free lifestyle assessment, and a summary of the preliminary research findings.

Participants were encouraged to bring a coworker whom they knew well, preferably the coworker they knew best. The participants completed a series of questionnaires during their first visit, and again 6 months later. Each battery of questionnaires took approximately 2 hr to complete. Participants completed a self-assessment (employee questionnaire) as well as an assessment of their coworker on the same items (coworker questionnaire).

RESULTS

The data were examined using bivariate scatter plots, scatter plots of the residuals, and expected normal probability plots. These showed that the assumptions of linearity and normality were basically met. The minimum and maximum values, means, standard deviations, and reliability coefficients for the scales of Altruism, SM, SD, and task performance are reported in Table 4.

Elimination of Potential Confound

Before testing the hypotheses, it was necessary to investigate a potential confounding variable for Hypotheses 2a and 2b. It is predicted that SD will predict agreement category such that those who are underestimators or who are in-agreement raters are more likely to have higher levels of SD than those that are overestimators. Because half the sample completed ratings on their peers before completing ratings on themselves, it is possible that there may be a significant difference on self-ratings of Altruism depending on whether self-ratings or peer ratings were completed first. For example, for those who rated themselves first (and who would not have an indication that peer appraisal would happen), the processes of SD may operate differently than for those who rated their peers prior to self-rating. If this turned out to be the case, it would be inappropriate according to the logic of Hypotheses 2a and 2b to include those who self-rated first into the analyses. To test for this potential confound, a *t* test was conducted. There was no significant difference between means on self-ratings of Altruism for those who rated themselves first ($M = 2.97$) compared to those who rated their peers first ($M = 3.02$), $t(397) = -1.23, p > .05$. Because no effect was found, all cases were included in the subsequent analyses.

Contextual Performance

Using agreement categories (underestimators, overestimators, and in-agreement raters) based on ratings of Altruism, a one-way analysis of variance (ANOVA) was conducted to see if there were any differences across these categories in terms of the OD subscale of SM. There was a significant difference between means, $F(2, 391) = 3.48, p < .05$. Contrary to Hypothesis 1a, in-agreement raters were not characterized by higher levels of SM. In fact, underestimators ($M = 2.31$) were significantly higher in OD than overestimators ($M = 2.16$), but not higher than in-agreement raters ($M = 2.23$). Another ANOVA was performed to see if the agreement categories differed on levels of SD. There was a significant difference between means, $F(2, 394) = 3.52, p < .05$. Post hoc comparisons using the Tukey HSD model revealed that underestimators are significantly higher in SD ($M = 15.75$) than overestimators

TABLE 4
 Minimum and Maximum Values, Means, and
 Standard Deviations for Altruism, Task Performance,
 Social Desirability, and Self-Monitoring

<i>Scale</i>	<i>N</i>	<i>Number of Items</i>	<i>Minimum</i>	<i>Maximum</i>	<i>M</i>	<i>SD</i>	<i>Coefficient α</i>
Altruism Self	400	6	1.67	4.00	3.02	.455	.75
Altruism Coworker	401	6	1.17	4.00	2.93	.490	.79
Task Performance Self	366	5	2.40	5.00	3.95	.554	.67
Task Performance Coworker	368	5	1.00	5.00	3.83	.646	.76
Social Desirability	407	33	2.00	31.00	14.71	5.96	.83
Self-Monitoring	404	25	1.28	3.28	2.40	.281	.75

($M = 14.08$), but not higher in SD than in-agreement raters ($M = 14.16$). Thus, although Hypothesis 2a was not supported, there was support for Hypothesis 1a.

To investigate the possibility that the method used to produce the agreement categories influenced the results, additional ANOVAs were conducted using different cut points to create the agreement categories. New agreement categories were created such that overestimators were characterized by a Z_{SAIt} greater than one standard deviation above the Z_{PAIt} ($N = 88$), underestimators by a Z_{SAIt} less than one standard deviation below the Z_{PAIt} ($N = 89$), and in-agreement raters by a Z_{SAIt} one standard deviation above or below the Z_{PAIt} ($N = 220$). The results of the ANOVA conducted on these categories replicated those based on the original categories. There were significant differences between the groups in terms of OD, $F(2, 391) = 3.82, p < .05$, as well as SD, $F(2, 394) = 3.48, p < .05$. Tukey's post hoc comparisons reveal again that underestimators were significantly higher in OD ($M = 2.38$) and SD ($M = 15.91$) than overestimators ($M_{OD} = 2.22; M_{SD} = 13.56$). When the cutoff for over- and underraters is left at one standard deviation and the bandwidth for in-agreement raters is changed to one-half a standard deviation, the results remain stable (i.e., underraters are significantly higher in SD than overerraters). The results do not remain stable when the cutoff points exceed one and a half standard deviations. Although the pattern of group means remains consistent for overestimators ($M_{OD} = 2.22; M_{SD} = 14.08$), underestimators ($M_{OD} = 2.37; M_{SD} = 16.46$), and in-agreement raters ($M_{OD} = 2.32; M_{SD} = 14.76$), the differences are not significant due to the heterogeneous group sizes in the agreement categories (for SD overestimators = 48, underestimators = 48, and in-agreement raters = 220; for OD overestimators = 48, underestimators = 47, and in-agreement raters = 219).

In accordance with Cohen and Cohen's (1983) recommendations for dealing with difference or change data, a hierarchical regression was conducted. Using self-ratings as the dependent variable, coworker ratings were entered as the covariate the first step of the regression. This remaining residual represents self-ratings with the effect of coworker ratings removed; that is, the extent of over- versus underestimating. At the second step of the regression procedure, OD and SD were entered together. Evidence for the replication of the post hoc analysis of the previous ANOVA would be found if both OD and SD had significant and negative beta weights. This is in fact what the analysis revealed: $\beta_{OD} = -.13, p < .05$; $\beta_{SD} = -.14, p < .005$. In other words, there was a significant association between underestimation and OD and underestimation and SD. These results corroborate the previous ANOVA results.

Task Performance

These analyses were repeated using the rating category scheme based on task performance. Unlike with contextual performance, there were no significant dif-

TABLE 5
Convergence of Self-Peer Ratings Across Measures Varying in Verifiability

Measure	Correlation Between Self and Peer Ratings (r)	N	r'
Attendance ^a	.426**	343	.455
Promptness ^a	.359**	342	.376
Quantity of work ^a	.211**	342	.214
Quality of work ^a	.148**	342	.149
Altruism ^b	.135*	342	.136
Overall performance ^a	.130*	343	.131

Note. r' refers to Fisher's r to z transformation, which is necessary in order to test the difference between two correlations. The term r' is used to avoid confusion with the standard normal deviate (Howell, 1992).

^aThese measures are at the item level. Overall performance is an item, not the task performance scale constructed from all the items that was used to create agreement categories for task performance.

^bAltruism is a scale constructed from the mean of seven items.

* $p < .001$. ** $p < .05$.

ferences across categories in terms of OD, $F(2, 338) = 2.02, p > .05$, or SD, $F(2, 340) = 2.3, p > .05$. However, in terms of SD, the pattern of means were similar to those for contextual performance. Underestimators were higher in SD ($M = 15.22$) and OD ($M = 2.29$) than overestimators ($M_{SD} = 13.66; M_{OD} = 2.23$); however, there were no significant differences between these means.

As with contextual performance, new agreement categories were created for task performance using the criterion of one standard deviation (overestimators $> 1 SD +$ standardized peer rating of task performance, underestimators $< 1 SD$, and in-agreement raters within $\pm 1 SD$) to check the stability of the results. The nonsignificant results were replicated for OD, $F(2, 340) = 1.97, p > .05$, and SD, $F(2, 338) = .63, p > .05$. However, the pattern of means for SD remained consistent for over ($M = 13.49$), under ($M = 15.27$), and in-agreement raters ($M = 14.91$).

The same regression procedure used for contextual performance to assess agreement on a continuous scale was conducted using self-ratings of task performance as the dependent variable, partialling out the effect of coworker ratings of task performance. When OD and SD are entered as the predictors, no relation of OD with underestimation or overestimation ($\beta = -.05, p > .05$) was found. However, SD was associated with underestimation ($\beta = -.12, p < .05$). Taking these results into account, there is support for Hypothesis 2b but no support for 1b.

Verifiability of Performance Dimension

To test the third hypothesis, correlations between self-ratings and peer ratings across the performance dimensions were computed. As Table 5 indicates, ratings of attendance tended to converge the most whereas ratings of Altruism and overall

performance converged the least. The convergence on ratings of attendance was significantly higher than convergence on ratings of quantity of work ($z_{\text{obt}} = 3.15 > z_{.025} = \pm 1.96, p < .05$). The convergence of quantity of work was not significantly higher than the convergence of quality of work ($z_{\text{obt}} = .848 < z_{.025} = \pm 1.96, p > .05$). Convergence for quality of work was neither significantly higher than convergence on Altruism ($z_{\text{obt}} = .175 < z_{.025} = \pm 1.96, p > .05$), nor was convergence for Altruism significantly higher than convergence for overall performance ($z_{\text{obt}} = .067 < z_{.025} = \pm 1.96, p > .05$).

DISCUSSION

The purpose of this study was to determine whether SM and SD were associated with rater agreement and different types of rater disagreement (overestimators and underestimators). When these rating categories were formed using ratings of contextual performance, there was a significant relation between under estimation and the OD subscale of SM as well as SD. Underraters were characterized by higher levels of OD and SD than overestimators. Although this difference between under- and overestimators was significant for rating categories based on contextual performance, the difference failed to reach significance for task performance. Although it is interesting to note that the means for SD across agreement categories for task performance showed the same general ranking as for contextual performance (underestimators > in-agreement raters > overestimators), the differences between means, although not significant, are notable because the magnitude of the differences between overestimators and underestimators on SD were similar for task ($d = 1.55$) and contextual ($d = 1.67$) performance.

This support for Hypothesis 2a suggests that in this context using the MCSD, it was socially desirable to be modest about self-ratings. Those individuals higher in SD (i.e., higher need for approval) may be motivated to rate themselves modestly, which would explain the higher scores on SD for underestimators. Moreover, overraters scored lower on SD. This may indicate that overraters were not as compelled to seek social approval, and did not mind representing themselves in a more positive manner than their peers. Further, Crowne (1979) remarked that the need for approval (as assessed by the MCSD) can also be thought of as the need to avoid disapproval. In this study, those who were higher on the MCSD may have been motivated to underrate in order to avoid disapproval upon the comparison of their ratings with those of their peers. However, it is important to note that the conceptualization of SD is dependent on which scale is used. Future research on individual differences and rating congruence must distinguish between measures of SD because those measures may lead to different results. For instance, the Balanced Inventory of Desirable Responding (Paulhus, 1984) taps self-deceptive enhancement and impression management, and research utilizing this measure may find that overraters may

be distinguished by higher levels of SD because the self-ratings of one high on this scale may be positively enhanced when compared to peer ratings. Therefore, the results of this study clearly point to the need for approval as an individual difference variable that affects S–O rating convergence; the impact of one's desire to make him or herself look good as it impacts convergence is left to be examined.

Contrary to Hypotheses 1a and 1b, in-agreement raters were not characterized by a higher degree of SM as assessed by OD; rather, underestimators were. In other words, higher self-monitors were presenting themselves in a less desirable light than their peers were presenting them. This goes against what Church (1997) found—that people higher in SM tended to obtain more convergence in S–O ratings. However, the sample in the Church study consisted of managers; our sample consisted of operating level employees with the majority having less than 2 years of college education. Moreover, Church used an overall measure of SM, whereas the subscale of OD was used here.

Taken together, these differences may account for the discrepant findings between this study and that of Church (1997). For example, an examination of the items on OD shows that this subscale is more reflective of pleasing others, acting according to others' expectations, and generally changing behavior to be liked or accepted. Lennox and Wolfe (1984) noted how OD taps attention to social comparison information, and is in fact what links it to the overarching construct of SM. This notion is similar to the logic behind how SD is operating on convergence, and the scales are indeed correlated ($r = .33, p < .05$). However, OD is conceptually distinct from SD because it seems to be assessing the extent to which one *acts* in order to gain acceptance rather than one's *need* for approval. Therefore, the results of this study indicate that the construct of SM may operate differently on convergence for different samples and different measures of SM.

There was limited support for Hypothesis 3. Attendance—a verifiable dimension of job performance—obtained the highest degree of self–peer rating convergence and was significantly higher than any of the other convergence ratings. As noted earlier, this may be due to the equal access of self and peers to unambiguous information (i.e., whether or not someone is present at work is obvious to both raters). Although the magnitude of the convergence values are in the expected direction (quantity > quality, quality > Altruism, quantity > Altruism), these values are not significantly different from one another. These findings support those of Conway (1996), who found no difference in interrater reliabilities for task versus contextual performance. Moreover, the convergence values are lower than what we would expect given Harris and Schaubroeck's (1988) meta-analytic work on the convergence between self- and peer ratings (.36). Although these values fall well within the range of their 90% confidence intervals, there may be characteristics of the sample that caused these values to be low. For instance, this sample included less educated participants who may not be as skilled at rating behavior.

Also, the results of the meta-analysis included within-occupation performance measures whereas in this study, a general performance measure was used to assess general task performance across many types of jobs.

Practical Implications

The finding that underestimators had higher levels of OD and SD than overestimators and in-agreement raters is interesting, especially because participants were asked to bring a coworker whom they knew well to participate in the study. This has implications for how raters are selected in multirater systems. A common assumption is that when ratees choose their peer raters, this is essentially stacking the odds in the ratee's favor. However, the findings here indicate that not only is there variance in the amount of convergence, but that those who tend to underrate themselves relative to peer ratings may be doing so out of modesty. This modesty may be a function of the target having chosen their peer raters. Thus, when ratees choose their raters, and have a high need for approval, there may be a tendency for a modesty effect rather than a positive bias—a finding contrary to the conventional wisdom in this area.

Another implication of this study is the degree of convergence practitioners of 360° feedback can expect. The nature of the performance dimension being rated will affect the upper limits of how much S-O convergence is possible. As the criterion space of job performance is expanded to include contextual performance, it will become increasingly important how rating discrepancies of citizenship behavior are interpreted. In addition, it may be necessary to take into account the verifiability of the performance dimension being rated because this influences the degree of convergence obtained. When using contextual performance measures in the 360° feedback system, discrepancies in these ratings should be interpreted after the effect for verifiability has already been accounted for. The meaning attached to discrepancies in ratings of citizenship behaviors must be interpreted in the context of how much convergence we should expect given the ambiguous nature of the construct of citizenship, especially as it compares to more objective measures of task performance. Although meta-analytic work will better inform these expectations, this study is a step in this direction such that ratings of citizenship do not tend to converge as much as ratings of attendance or quantity of work.

It is important to note that the finding that people higher in SD were more likely to be underraters was only significant for ratings of contextual performance. This is related to the investigation of the upper limits of rating convergence on citizenship. Although the general trend for underestimators to have higher SD remained consistent for task performance, the differences were not significant across rating categories. This may imply that, in addition to choosing peer raters, modesty may be driven by the nature of the performance construct being rated. Individuals may

be more likely to give lower ratings relative to their peers on measures of contextual performance, which would consequently lower the amount of S-O rating convergence capable of being obtained.

Limitations

The most salient limitation to this study is that data were not obtained in an actual 360° feedback setting. The data were collected within a large research effort to examine variables of employee lifestyle, physical and mental health, and organizational outcomes. Although self- and peer responses were collected on several of these variables, the rating situation did not adequately simulate a multirater system. Thus, the specific motivational and cognitive influences driving the rating processes may have affected the ratings differently than they might have if the data were collected in an actual 360° feedback situation.

Moreover, there were limitations in the method used to capture the construct of convergence. This method involved categorizing a variable that is conceptualized as continuous. This is considered to be a thorny issue, and researchers have measured convergence using a variety of methods including interrater reliability (Fleenor et al., 1996), difference scores, Pearson correlations between the average self-score and the corresponding average others' ratings (Church, 1997), the regression method using the difference of weights and intercepts (Brutus, Fleenor, & McCauley, 1996), the polynomial regression technique (Brutus, Fleenor, & Tisak, 1996; Edwards, 1994a; Edwards, 1994b), and the method used presently, categories of convergence (Atwater & Yammarino 1992, 1997; Van Velsor et al., 1993). As Brutus, Fleenor, and Tisak (1996) noted, there are shortcomings inherent in all these methods, including the categorization technique. However, the convergence category approach has a practical appeal to those who utilize multirater systems. This simplicity in the measurement of convergence makes it more intuitive for managers to understand. Contrast this with the complexities in explaining for example, the mechanisms of why difference scores tend to be unreliable and attenuate any relation that may exist between predictor and criterion, and the clarity of the rating category approach becomes evident. Therefore, although there are drawbacks in capturing convergence by constructing agreement categories, this method is clearly superior to the other options for practical utility reasons.

Directions for Future Research

Given the issues that emerged from this study, three avenues for future research become clear. The first is establishing a conceptual framework that prescribes when convergence should/should not be expected and when it should/should not be de-

sired. The delineation of the situational complexities of rating convergence will help inform practice. S-O agreement may mean different things across organizational settings, and across performance dimensions. This is especially true as ratings of contextual performance become more common in multirater systems. For example, lack of convergence on ratings of contextual performance may be more an indication of the variance in implicit theories about what constitutes contextual performance than rater or ratee accuracy. Therefore, instead of using convergence as a catchall proxy for the quality of ratings, practitioners and researchers alike must gain a better understanding of the driving mechanisms of convergence across performance domains—especially as convergence is increasingly used as a predictor of organizational outcomes.

Secondly, the stability of convergence as an individual difference variable as well as an interindividual difference variable (because it takes two individuals to obtain convergence) must be explored to better understand it both as a predictor and a criterion. Understanding what convergence really is and what accounts for variance in it is essential before any recommendations are given regarding when and where we should expect it to occur. This will entail further investigations into the individual, dyadic, group, and organizational characteristics that impact the degree of rater agreement.

Thirdly, research is sorely needed in how discrepant feedback should be interpreted and dealt with. As the conceptual framework for convergence is developed and refined, we may gain a better understanding of what to do with discrepant feedback once it is received. For example, it would be useful to know how to minimize defensiveness and threats to self-efficacy caused by rating discrepancies, and use them as positive levers for change. The feedback process is truly where the rubber meets the road regarding rating convergence, and research must begin to focus on the impact convergence has on these processes.

Continued research in rating convergence may illuminate more contingencies than answers. However, given the growing prevalence of 360° feedback systems, issues surrounding rater agreement will not likely fade into the background. Practice must feed back to research and *visa versa* so that the utility and quality of multirater systems can continue to be enhanced.

ACKNOWLEDGMENTS

This work was supported by National Institute of Mental Health Grant # R03 MH 50230-02.

An earlier version of this article was presented at the annual meeting of the American Evaluation Association in Chicago, November 1998.

We thank Jane Davidson, Rudolph Sanchez, and Kathryn Paget for their comments on earlier versions of this article. We would also like to acknowledge the anonymous reviewers of this article for their helpful feedback.

REFERENCES

- Ashford, S. (1989). Self-assessments in organizations: A literature review and integrative model. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior* (Vol. 11, pp. 133–174). Greenwich, CT: JAI.
- Atwater, L., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self-and follower ratings of leadership. *Personnel Psychology, 48*, 35–59.
- Atwater, L., & Yammarino, F. J. (1992). Does self–other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Personnel Psychology, 45*, 141–164.
- Atwater, L., & Yammarino, F. J. (1997). Antecedents and consequences of self–other rating agreement: A review and model. In G. Ferris (Ed.), *Research in personnel and human resources management* (pp. 121–174). Greenwich, CT: JAI.
- Bass, B., & Yammarino, F. J. (1991). Congruence of self and others' leadership ratings of naval officers for understanding successful performance. *Applied Psychology: An International Review, 40*, 437–454.
- Borman, W. (1974). The rating of individuals in organizations: An alternative approach. *Organizational Behavior and Human Performance, 12*, 105–124.
- Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmit & W. Borman (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco: Jossey-Bass.
- Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*, 99–109.
- Bozeman, D. P. (1997). Interrater agreement in multi-source performance appraisal: A commentary. *Journal of Organizational Behavior, 18*, 313–316.
- Briggs, S. R., Cheek, J. M., & Buss, A. H. (1980). An analysis of the self-monitoring scale. *Journal of Personality and Social Psychology, 38*, 679–686.
- Brutus, S., Fleenor, J. W., & McCauley, C. D. (1996, April). *Self–other rating discrepancy in 360-degree feedback: An investigation of demographic and personality predictors*. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, San Diego.
- Brutus, S., Fleenor, J. W., & Tisak, J. (1996, April). *Measuring congruence in 360-degree feedback research*. Poster session presented at the meeting of the Society for Industrial and Organizational Psychology, San Diego.
- Carver, S. C., & Scheier, M. F. (1992). *Perspectives on personality* (2nd ed.). Boston: Allyn & Bacon.
- Church, A. H. (1997). Managerial self-awareness in high-performing individuals in organizations. *Journal of Applied Psychology, 82*, 281–292.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Conway, J. M. (1996). Additional construct validity evidence for the task/contextual performance distinction. *Human Performance, 9*, 309–329.
- Crowne, D. P., & Marlow, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349–354.
- Crowne, D. P., & Marlow, D. (1964). *The approval motive*. New York: Wiley.
- Donaldson, S. I., Ensher, E. A., & Grant-Vallone, E. J. (2000). Longitudinal examination of mentoring relationships on organizational commitment and citizenship behavior. *Journal of Career Development, 26*, 233–249.
- Edwards, J. R. (1994a). Regression analysis as an alternative to difference scores. *Journal of Management, 20*, 683–689.
- Edwards, J. R. (1994b). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Process, 58*, 51–100.

- Fleenor, J. W., Brutus, S., & McCauley, C. D. (1996, August). *Does self-other rating agreement moderate the relationship between personality and leader effectiveness?* Paper presented at the meeting of the American Psychological Association, Toronto, Ontario.
- Furnham, A., & Stringfield, P. (1994). Congruence of self and subordinate ratings of managerial practices as a correlate of supervisor evaluation. *Journal of Occupational and Organizational Psychology*, *67*, 57-67.
- Furnham, A., & Stringfield, P. (1998). Congruence in job-performance ratings: A study of 360 feedback examining self, manager, peers, and consultant ratings. *Human Relations*, *51*, 517-530.
- Gabrenya, W. K., Jr., & Arkin, R. M. (1980). Factor structure and factor correlates of the self-monitoring scale. *Personality and Social Psychology Bulletin*, *6*, 13-22.
- Harris, M. H., & Schaubroeck, J. (1988). A meta-analysis of self-supervisory, self-peer, and peer-supervisor ratings. *Personnel Psychology*, *41*, 43-62.
- Howell, D. C. (1992). *Statistical methods for psychology* (3rd ed.). Belmont, CA: Duxbury.
- Lennox, R. D., & Wolfe, R. N. (1984). Revision of the self-monitoring scale. *Journal of Personality and Social Psychology*, *46*, 1349-1364.
- Mabe, P. M., & West, S. G. (1982). Validity of self-evaluations of ability: A review and mental analysis. *Journal of Applied Psychology*, *67*, 280-296.
- Mersman, J. L., & Shultz, K. S. (1998). Individual differences in the ability to fake on personality measures. *Personality and Individual Differences*, *24*, 217-227.
- Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.
- Nilsen, D., & Campbell, D. P. (1993). Self-observer rating discrepancies: Once an overrater, always an overrater? *Human Resource Management*, *32*, 265-281.
- Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA: Heath and Company.
- Organ, D. W. (1997). Organizational citizenship behavior: It's construct clean-up time. *Human Performance*, *10*, 85-97.
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*, 598-609.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaves, & L. S. Wrightsman's (Eds.), *Measures of personality and social psychological attitudes* (pp. 12-59). San Diego, CA: Academic.
- 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*, 813-820.
- Smith, C. A., Organ, D. W., & Near, J.P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology*, *68*, 653-663.
- Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology*, *30*, 526-537.
- Vance, R. J., MacCallum, R. C., Coovert, M. D., & Hedge, J. W. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. *Journal of Applied Psychology*, *73*, 74-80.
- Van Velsor, E., Taylor, S., & Leslie, J. (1993). An examination of the relationships between self-perception accuracy, self-awareness, gender and leader effectiveness. *Human Resource Management*, *32*, 249-264.
- Wicklund, R. (1975). Objective self-awareness. In L. Berkowitz, (Ed.), *Advances in experimental social psychology* (Vol. 8, pp. 233-275). New York: Academic.
- Williams, J. R., & Levy, P. E. (1992). The effects of perceived system knowledge on the agreement between self-ratings and supervisor ratings. *Personnel Psychology*, *45*, 835-847.

- Wohlers, A. J., Hall, M., & London, M. (1993). Subordinates rating managers: Organizational and demographic correlates of self/subordinate agreement. *Journal of Occupational & Organizational Psychology*, *66*, 263–275.
- Wohlers, A. J., & London, M. (1989). Ratings of managerial characteristics: Evaluation difficulty, subordinate agreement, and self-awareness. *Personnel Psychology*, *42*, 235–261.
- Yammarino, F. J., & Atwater, L. E. (1993). Understanding self-perception accuracy: Implications for human resource management. *Human Resource Management*, *32*, 231–248.
- Yammarino, F. J., & Atwater, L. E. (1997). Do managers see themselves as others see them? Implications of self–other rating agreement for human resources management. *Organizational Dynamics*, *25*(4), 35–44.

Copyright of Human Performance is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.