THE EFFECTS OF INTRERRATER AND SELF-OTHER AGREEMENT ON PERFORMANCE IMPROVEMENT FOLLOWING UPWARD FEEDBACK

JEFF W. JOHNSON
Personnel Decisions Research Institutes, Inc.

KERRI L. FORSTI
University of Minnesota and Personnel Decisions Research Institutes, Inc.

This study investigated (a) whether performance improvement following upward feedback is related to self–other rating discrepancies, (b) how self-ratings change after feedback, and (c) whether agreement among raters influences performance improvement. Self-ratings and subordinate ratings were collected from 1,888 managers at 2 points in time 1 year apart. Using polynomial regression equations, we determined that managers who overrated themselves relative to how others rated them tended to improve their performance from 1 year to the next, and underraters tended to decline. This is consistent with what would be predicted by self-consistency theory (Korman, 1976). Self-ratings tended to decrease for overraters and increase for underraters, but this effect was not constant throughout the range of self-ratings. Agreement among subordinate raters was negatively related to performance improvement, but this is likely a reflection of differential regression to the mean.

Despite the current widespread practice of collecting performance ratings from subordinates to aid in the development of managers (i.e., upward feedback), there has been little research linking characteristics of upward feedback to behavior change. It is often assumed that discrepancies between self-ratings and subordinate ratings raise self-awareness, highlight gaps between goals and job performance, and suggest areas in need of improvement (London & Smither, 1995; Tornow, 1993). Users of upward feedback programs hope, if not expect, that managers will respond to these insights by actually taking steps to improve their performance.

An earlier version of this paper was presented at the 12th Annual Conference of the Society for Industrial and Organizational Psychology, Inc., St. Louis, MO, April, 1997.

We thank James W. Smither, Avraham N. Kluger, and the anonymous reviewers for their helpful comments.

Correspondence and requests for reprints should be addressed to Jeff W. Johnson, Personnel Decisions Research Institutes, 43 Main Street SE, Ste. 405, Minneapolis, MN 55414; jeffj@pdi-corp.com.
The purposes of the present study were to investigate (a) whether performance improvement following upward feedback is related to self–other rating discrepancies, (b) how self-ratings change after feedback, and (c) whether agreement among raters influences performance improvement. Smither et al. (1995) suggested that self-consistency theory (Korman, 1970, 1976) might explain the relationship between self–other agreement and performance improvement. Self-consistency theory, if applied to upward feedback, suggests that people desire congruence between their own and others' perceptions of their behavior. The prediction follows that the presence and direction of self–other rating discrepancies will be related to changes in self-evaluations and performance following feedback. London and Smither (1995) proposed that self–other agreement interacts with interrater agreement (e.g., among subordinates, between supervisors and peers) to influence a ratee's attention to and acceptance of feedback. The more the ratee attends to and accepts feedback, the more performance is expected to improve (Ilgen, Fisher, & Taylor, 1979).

Self–Other Agreement and Subsequent Performance

When managers receive feedback indicating that their subordinate ratings are lower than their self-ratings, they might begin to question whether their behavior is consonant with their self-image. According to self-consistency theory (Korman, 1970, 1976), these managers should be motivated to improve their performance, thus reducing the discrepancy between how they perceive themselves and how others perceive them and restoring a sense of cognitive balance. If feedback is consistent with self-perceptions, however, managers may be satisfied and not be motivated to improve, even if their current performance level is low (Korman, 1970, 1976). Similarly, managers whose subordinate ratings exceed their self-ratings would have little incentive to improve their performance.

In an investigation of performance improvement following upward feedback, Smither et al. (1995) studied a group of managers in a large corporation as an upward feedback program was being initiated. Managers were told that the program was part of an organizational change effort, that their participation was voluntary, and that the only purpose of the feedback was to guide their growth and development. Self- and subordinate ratings were collected at two points in time 6 months apart. Those managers whose Time 1 mean subordinate ratings were in the bottom third of the sample distribution (n = 80) tended to be the ones who most highly overrated themselves with respect to their subordinates' ratings. Although Smither et al. did not report the number of overraters in the low performer sample, the correlation between subordinate ratings and self-minus-subordinate ratings in the total sample at Time 1 was -.77. As a group, these low-performing overraters significantly improved their performance following upward feedback. Although some of this improvement can be explained by regression to the mean, the authors reported an effect size of about d = .25 for performance gains beyond those expected by regression effects. Further, follow-up on 31 of these initially low-performing managers after 1.5 and 2.5 years showed that performance gains were sustained over time (Reilly, Smither, & Vasiropoulos, 1996).

To test the prediction that low performers would not improve if they were in-agreement or underraters, Smither et al. (1995) selected those managers whose mean subordinate ratings were less than 4.00 on a 5-point scale, and whose self-ratings were equal to or lower than their subordinate ratings (n = 19). These managers did not improve their performance following feedback, but this result must be interpreted with caution because of the small number of managers in this subgroup. Overall, results from the low initial performers supported predictions based on self-consistency theory.

Atwater, Roush, & Fischthal (1995) investigated changes in leadership ratings as a function of initial self–other agreement. Their participants were 1,232 freshmen and 978 juniors in the U.S. Naval Academy, where juniors (leaders) were responsible for the training of freshmen (followers). Leaders completed self-ratings and received aggregated, anonymous ratings from their followers. Sixteen weeks after receiving feedback, leaders were evaluated again. Consistent with predictions, there was a significant interaction between time of rating and agreement category. Specifically, follower ratings of overraters increased significantly, and there was no significant change for underraters or in-agreement raters. The authors provided evidence that the results were not simply regression effects.

In the Atwater et al. (1995) study, as in others (e.g., Smithet et al., 1995; Van Velsor, Taylor, & Leslie, 1993), self–other agreement was strongly related to initial level of performance. That is, overraters tended to have the lowest initial performance and underraters the highest, with in-agreement raters in the middle. This relationship is not surprising, given that the mean others' rating is the index of initial performance and is also used to determine self–other agreement. When agreement is thus confounded with initial performance, alternative explanations of performance change are possible. In other words, managers who are overraters at the time of initial assessment might improve their performance after feedback because they were the lowest performers to begin with. There are many ways that providing relatively negative
feedback can cause people to improve their performance (see Johnson, Olson, & Courtney, 1996). For example, feedback provides people with information about what behaviors are expected and valued by the organization, what specific aspects of their performance are in need of improvement, and what opportunities are available to help improve their skills. This knowledge will help the poorest performers the most, because they are the ones who are most in need of improvement. If we assume that most people would like to be good performers, then in general, those who receive low ratings should be more motivated to improve their performance than those who receive high ratings. Because of the relationship between initial level of performance and self-other agreement, it is important to separate these variables when examining the influence of self-ratings on performance improvement. By looking at the small number of low initial performers who were underraters or in-agreement raters separately from the larger group of low initial performers, Smither et al. (1995) demonstrated support for the hypothesis that self-other rating discrepancies, not initial performance, explain reactions to upward feedback. Specifically, poor performers tended to improve after feedback, but not if their self-ratings were in agreement with or lower than their subordinate ratings. As Smither et al. acknowledged, however, the low power associated with this result (based on n = 19) points to the need for replication and further study.

Because a large sample was available for the present study, we were able to examine the effect of self-other agreement on performance improvement for a wide range of initial subordinate ratings. Based on self-consistency theory (Korman, 1970, 1976) and the research reviewed above, the following hypotheses were advanced:

**Hypothesis 1a:** The subordinate ratings of overraters will tend to increase following upward feedback, regardless of the level of initial subordinate ratings.

**Hypothesis 1b:** The subordinate ratings of underraters and in-agreement raters will tend to stay the same following upward feedback, regardless of the level of initial subordinate ratings.

**Self-Other Agreement and Subsequent Self-Evaluation**

When confronted with self-other rating discrepancies, it is reasonable to expect that managers would revise their self-evaluations in addition to (or instead of) changing their behavior. Underlying Korman's (1976) hypothesis that people are motivated to behave in ways that are consistent with self-perceptions is the assumption that others' evaluations are one determinant of those self-perceptions. In other words, self-perceived competence concerning a task is influenced by the extent to which others believe that we are competent and communicate that belief to us. Korman reviewed some evidence supporting this assumption (see also Shrauger & Rosenberg, 1970; Shrauger & Schoenemann, 1979).

In their study of U.S. Naval Academy leaders and followers, Atwater et al. (1995) examined postfeedback changes in self-ratings for overraters, underraters, and in-agreement raters. Consistent with their hypotheses and with predictions suggested by self-consistency theory (Korman, 1970, 1976), self-ratings decreased for overraters, increased for underraters, and did not change for in-agreement raters. The changes in self-ratings exceeded those that would be expected from regression to the mean. Based on these findings and self-consistency theory, the following hypothesis was advanced:

**Hypothesis 2:** Following upward feedback, self-ratings will tend to increase for underraters, decrease for overraters, and remain constant for in-agreement raters.

**Interrater Agreement**

Users of upward feedback hope that below average managers who hold unjustifiably favorable views of their own performance will become more self-aware after upward feedback and set goals to improve their performance (Tornow, 1993). Such a response cannot be expected, however, unless the rater accepts the feedback as an accurate portrayal of his or her performance (Atwater & Yammarino, 1997; Ilgen et al., 1979). According to attribution theory (Weiner, 1986), when subordinates' ratings differ from self-ratings, the rater will analyze possible reasons for those discrepancies. London and Smither (1995) suggested that disagreement among subordinates (i.e., high variability in their ratings) may cause a rater to dismiss the ratings as reflecting idiosyncrasies of the raters rather than accept them as reflections of his or her own behavior. If others' ratings agree closely with one another, however, the rater may pay more attention to self-other discrepancies and feel more pressure to be able to explain those discrepancies. If raters are more attune to and accepting of ratings when raters agree with one another, it follows that raters will tend to improve more with greater interrater agreement. Based on London and Smither's reasoning, the following hypothesis was advanced:
Hypothesis 3: Among managers who overrate themselves compared to subordinate ratings at Time 1, there will be a positive relationship between interrater agreement at Time 1 and performance improvement following feedback.

Method

Participants

Participants were members of a large accounting firm who participated in two administrations of an upward feedback program. Ratees were managers at three levels (managers, senior managers, and partners); raters were the direct reports of these managers. The upward feedback program was first implemented in 1994, but data from that administration were not available for this study. For 1995, a new consulting firm was chosen to process and report the feedback data, and changes were made in the content of the rating instrument. No changes were made between 1995 and 1996, the years for which data were available for this study. In 1995, 3,099 managers participated as ratees, all of whom were rated by at least three raters. Of these managers, 2,171 were also rated by at least three raters in 1996, and all analyses were conducted on this sample.

Of the 2,171 managers who participated in both 1995 and 1996, 1,903 (87.7%) also completed self-ratings in 1995, and 2,091 (96.3%) completed self-ratings in 1996. The average number of raters per rater in 1995 was 6.64 (SD = 2.10); the average number of raters per rater in 1996 was 7.34 (SD = 2.10). The sample was composed of 902 partners, 731 senior managers, and 514 managers (24 did not indicate their level). Managers were located in business units all over the United States, with the largest concentrations in the metropolitan New York (14.9%) and Midwest (14.8%) areas. Unfortunately, the number of managers who transferred to different areas between 1995 and 1996 was not known.

Instrument

The rating instrument was developed in 1994 to measure four broad performance categories identified as important by leaders of the organization. These categories were labeled (a) Leadership, (b) People Management (e.g., coaching, evaluation, counseling), (c) People Development, and (d) Communications. The original instrument contained 32 items. In 1995, the wording was changed for several items based on feedback from participants, and 4 new items were added to bring the number to 36. Each item was a behavioral statement, and raters were instructed to base their ratings on the extent to which the manager exhibited each behavior. Ratings were made on a 5-point scale where 1 = not at all and 5 = to a very great extent. A not applicable option was available for behaviors on which the rater had no information. Raters also had the opportunity to provide written comments. The content of the instrument remained unchanged for 1996.

Procedure

The purpose of the upward feedback program was to assist in the development of managers by pointing out their relative strengths and weaknesses as seen by their subordinates. Upward feedback results were not explicitly tied to raises or promotions. Each ratee's supervisor received a copy of that individual's feedback report, however, so it is possible that some ratees made arrangements with their supervisors to base some outcome (e.g., bonuses) on the results of the feedback. In other words, it is possible that some ratees were held more accountable for performance improvement than were others.

All employees at the level of manager or above were required to participate as ratees in the upward feedback program. In May of 1995, each manager was instructed to distribute the feedback instrument to as many as 10 of his or her subordinates. Some managers distributed more than 10 instruments, however, because 33 managers in 1995 were rated by more than 10 raters (1.5%), and 123 managers in 1996 (5.7%) were rated by more than 10 raters. The maximum number of raters was 18 in 1995 and 16 in 1996. Raters had 2 weeks to complete and return the instrument. Completed surveys were mailed directly to an independent consulting firm for data compilation and the development of feedback reports. Ratees followed the same procedure after completing their self-ratings.

Feedback reports were provided only to managers who were rated by at least three subordinates. Note that in 1995 all participants met this criterion. Each feedback report contained the following information:

1. Interpretive guidelines. This was an explanation of each section of the report and how the results should be interpreted.
2. Category and overall feedback results. This section displayed results at the performance category and overall level. Performance category scores were calculated by computing the mean of all items within a category. The overall score was the mean score across all items. Scores were displayed for the self-rating, the mean rating from subordinates, and the mean rating for all participants at the same organizational level as the individual. Scores were presented numerically and in bar graph
form. The percentage of raters who chose each response option across all items in the category was also displayed with each score.

3. Item-level results. This section displayed the same information as was presented at the category level, but results were at the individual item level. In addition to response option percentages, however, the actual number of raters choosing each response option was presented.

4. Comments. This section reported comments transcribed verbatim from the comments section of the rating instrument.

5. Suggestions for using feedback. This section provided suggestions for what managers should do as their next step. Specifically, managers were told (a) to discuss the results with someone they trust and respect; (b) where to find tips and resources to help them improve in different areas; (c) to develop a plan for improvement; (d) to hold a meeting with the people who provided feedback, for the purpose of sharing plans for improvement and asking for additional input; and (e) to invite ongoing feedback on an informal basis.

The same procedure was followed in 1996, but all managers did not participate at the same time, as they did in 1995. Managers were allowed to participate in the upward feedback program at any time between May and December of 1996. Therefore, at least 1 year but no more than a year and a half passed between administration in 1995 and administration in 1996.

Analyses

Overall score. Mean scores for each performance category were computed, but they were so highly intercorrelated (mean \( r \) for subordinate ratings = .85; mean \( r \) for self-ratings = .78) that it was decided that all analyses would be based on the overall measure (i.e., the average rating across all 36 items). Ratings from subordinates reporting to the same manager were averaged to create a single score for each item. The overall score was computed as the mean of these item scores. For both 1995 and 1996, Cronbach's alpha for the overall measure was .98 for mean subordinate ratings and .95 for self-ratings.

Because of the not applicable option on the response scale, the overall score was not necessarily based on all 36 items for each ratee. If all raters rated an item as not applicable, no score was available for that item when the overall score was calculated. Because the number of not applicable responses may change from one year to the next (e.g., as subordinates become more familiar with the manager's performance; Smith et al., 1995), this could affect the meaning of the overall score from year to year. In 1995, the mean number of items on which the overall subordinate score was based was 35.93, and 95.9% of the scores were based on all 36 items. In 1996, the mean number of items on which the overall subordinate score was based was 35.96 and 97.8% of the scores were based on all 36 items. We did not consider these differences to be large enough to affect the meaning of the overall score from year to year.

Interrater reliability and agreement. To evaluate the appropriateness of averaging subordinate ratings together, interrater reliability was determined by comparing the within-ratee variance to the between-ratee variance using a one-way ANOVA. The estimate of reliability for the overall score was calculated as (MSbetween - MSwithin)/MSbetween, which is ICC(1, k) from Shrout and Fleiss (1979). The reliability estimates were .60 for 1995 and .59 for 1996.

Interrater agreement was assessed for each item score using the \( r_{wg} \) index developed by James, Demaree, and Wolf (1984). This index reflects the extent to which individuals within the same setting agree on the rating of a stimulus compared to the agreement that would be expected by chance. Because upward feedback data are typically negatively skewed, the null distribution used for calculation of the index was that suggested for a slight negative skew (James et al., 1984). The average \( r_{wg} \) across items was calculated for each ratee. The mean across rates of these averages was .51 (SD = .17) for 1995 and .53 (SD = .17) for 1996. These results indicate that it was appropriate to aggregate individual ratings to the level of the ratee. Note, however, that there was also enough variability in interrater agreement that we were able to test our hypothesis regarding the effect of interrater agreement on performance improvement following feedback.

1994 participants. Although item-level data were not available for 1994, it was known that 1,891 (87.1%) of the managers who participated in 1995 and 1996 had participated in the first wave of the feedback program. A series of chi-square tests indicated that managers who participated in 1994 did not differ from managers who did not participate in 1994 in terms of their work location, but did differ in terms of their organizational level. As would be expected, those who did not participate in 1994 were more likely to be at lower levels of the organization, probably because many had just recently been promoted into a management position. In terms of 1995 subordinate ratings, managers who participated in 1994 \( (M = 3.98, SD = 0.37) \) did not differ from those who did not \( (M = 3.95, SD = 0.42; t = 0.85, n.s.) \). Managers who participated in 1994 did have significantly higher 1995 self-ratings \( (M = 3.94, SD = 0.43) \), however, than did managers who did not participate in 1994 \( (M = 3.84, SD = 0.46; t = 3.01, p < .01) \).
Changes in self- and subordinate ratings over time as a function of self-other agreement. Previous studies have used difference scores to operationalize self–other agreement and changes in ratings over time (e.g., Atwater et al., 1995; Hazucha, Hezlett, & Schneider, 1993; Smith, et al., 1995). Difference scores suffer from several methodological problems (Edwards, 1994, 1995; Brutus, Fleenor, & Taylor, 1996), not the least of which is the confounding of effects of the variables composing the difference score. An additional difficulty with difference scores for the present study is that both the independent variable and the dependent variable would be difference scores, and both difference scores would share components. For example, if performance improvement were operationalized as the difference between 1996 subordinate ratings and 1995 subordinate ratings, and self–other agreement were operationalized as the difference between 1995 self-ratings and 1995 subordinate ratings, both variables would include 1995 subordinate ratings as a component. This creates a complex relationship between the independent variable and the dependent variable that is very difficult to interpret.

Polynomial regression has been shown to avoid the problems associated with difference scores while allowing direct tests of the relationships represented by difference scores (Edwards, 1994, 1995; Edwards & Parry, 1993). The following equation can be used to model a variety of response surfaces:

\[ Z = b_0 + b_1X + b_2Y + b_3X^2 + b_4XY + b_5Y^2 + e \]  \hspace{1cm} (1)

For example, when investigating the effect of self–other agreement on changes in subordinate ratings, \( Z = 1996 \) subordinate rating, \( X = 1995 \) subordinate rating, and \( Y = 1995 \) self-rating. Including \( X \) in the equation captures the effect of 1995 subordinate rating as a covariate when assessing change in subordinate ratings, and as a comparison standard for 1995 self-ratings. Supplementing \( X \) and \( Y \) with curvilinear and interaction terms allows for tests of effects that depend on initial levels of \( X \) and \( Y \) (as suggested by Hypotheses 1a and 1b).

When investigating the effect of self–other agreement on changes in self-ratings, \( Z = 1996 \) self-rating, and \( X \) and \( Y \) are defined in the same way as above. In this case, \( Y \) captures the effect of 1995 self-rating as a covariate when assessing change in self-ratings, and as a comparison standard for 1995 subordinate ratings.

Regression to the mean. Part of our hypotheses is that subordinate or self-ratings will increase or decrease from year to year, but the influence of regression to the mean makes polynomial regression equations difficult to interpret in terms of whether scores actually increased or decreased. Regression to the mean predicts that scores below the mean in 1995 will increase in 1996, and scores above the mean in 1995 will decrease in 1996. Therefore, it is difficult to tell whether the predicted score given by a polynomial regression equation is beyond what would be expected by regression to the mean. To enhance interpretation, a predicted 1996 score based solely on regression to the mean is included whenever example data showing predicted 1996 scores based on a polynomial regression equation are presented.

Previous studies have used the correlation between Time 1 scores and Time 2 scores as an estimate of test-retest reliability to be used in a bivariate regression equation to predict an individual’s rating at Time 2, given the rating at Time 1 (e.g. Hazucha et al., 1993; Smith et al., 1995; Walker, 1997). Note, however, that the correlation between scores in two administrations of a feedback program is a very conservative estimate of reliability because some managers are likely to have changed their performance more than others as a result of feedback. If the estimate of reliability is too low, some real change in performance will be attributed to regression to the mean. Therefore, initially low performers could appear to have decreased their performance if the increase in scores is not as large as was predicted by the conservative estimate of reliability. Similarly, initially high performers who decrease their performance may appear to improve if regression to the mean is overestimated.

The correlation between subordinate scores in 1995 and subordinate scores in 1996 \((r = .51)\) is clearly an underestimate of test-retest reliability (i.e., the coefficient of stability). To assess test-retest reliability, the interval between measurements should be short enough that maturation or historical changes would not be expected to affect scores at Time 2 (Crocker & Algina, 1986). The feedback intervention in this study was expected to influence scores at Time 2, and the hypotheses advanced suggest that the feedback would not have the same effect on each participant. Furthermore, changes in performance independent of the feedback intervention would be expected in the 12- to 18-month interval between measurements. Therefore, a more appropriate dependent variable is a standardized residual that is based on a more realistic estimate of the coefficient of stability. Unfortunately, there is little research in this area to guide the choice of an estimate. In a meta-analysis, Viswesvaran, Ones, and Schmidt (1996) reported a weighted mean coefficient of stability for supervisor ratings of overall job performance of .81. This was for supervisor ratings rather than subordinate ratings, however, and only included studies in which the same rater was used at both points in time. In the present study, scores were mean ratings across several subordinates, and the ratings were not necessarily provided by all the same subordinates at each point in time. The study by Smith et al. (1995)
had characteristics similar to the present study, and they used a Time 1-Time 2 (6-month interval) correlation of .62 as a conservative estimate of stability. They used this estimate again in Reilly et al. (1996). We decided on an arbitrary but reasonable coefficient of stability estimate of .65. This is about midway between the lower bound estimate of .51 and Viswesvaran et al.’s estimate of .81, which may be considered an upper bound estimate. The predicted 1996 subordinate rating based solely on regression to the mean was calculated using the following formula:

\[ Z_{reg} = 0.65X \left( \frac{SD_z}{SD_X} \right) \]  

(2)

where \( X \) is the 1995 subordinate rating, \( SD_X \) is the standard deviation of the 1995 subordinate rating, and \( SD_z \) is the standard deviation of the 1996 subordinate rating.

The same logic applies to the change in self-ratings. The feedback intervention and our hypotheses suggest that the correlation between 1995 self-ratings and 1996 self-ratings (\( r = .60 \)) is too low for a coefficient of stability. The Viswesvaran et al. (1996) estimate of .81 for supervisor ratings is closer to what would be expected for self-ratings, so the coefficient of stability estimate for self-ratings was set at .80. The predicted 1996 self-rating based solely on regression to the mean was created using this estimate of reliability, in a similar manner as the predicted 1996 subordinate rating.

**Moderating effect of interrater agreement.** Hypothesis 3 predicted that interrater agreement among subordinates would moderate the relationship between self- and subordinate agreement and the change in subordinate ratings. According to Edwards (1994), the moderating effect of interrater agreement can be tested by hierarchically testing the product of interrater agreement and each term in the original model, after controlling for interrater agreement and the terms in the original model. A significant increase in \( R^2 \) suggests a moderating effect for interrater agreement. If interrater agreement intensifies the effects of self- and subordinate agreement on the change in subordinate ratings, the regression coefficients on the new terms will be in the same direction as those for the corresponding terms in the original model.

**Results**

**Change in Subordinate Ratings**

To avoid the possible influence of outliers on the results, cases were excluded if the score on 1995 subordinate rating, 1996 subordinate rating, or 1995 self-rating was greater than 3.5 standard deviations from the mean. This eliminated 15 cases. Means, standard deviations, and intercorrelations for all study variables after the elimination of outliers are presented in Table 1. Overall, there was a slight increase in subordinate ratings from 1995 to 1996, \( t(1,887) = 2.02, p < .05 \). Because of the low correlation between 1995 self- and subordinate ratings (.18), multicollinearity was not a problem for our regression analyses.

To maintain the interpretability of the regression coefficient for the interaction between \( X \) and \( Y \) (\( XY' \)), 1995 subordinate rating and 1995 self-rating were centered around the point halfway between their means (3.95495) before entering them into the regression analyses (Edwards, 1994). 1996 subordinate rating was also centered around this value so it could be more easily compared to the 1995 variables.

Hypothesis 1 predicted that overraters would tend to improve from 1995 to 1996, while in-agreement and underraters would tend to stay the same. This suggests positive regression coefficients for \( X \) (1995 subordinate rating) and \( Y \) (1995 self-rating), and a significant interaction between \( X \) and \( Y \). A positive regression coefficient for \( X \) would indicate that 1995 subordinate ratings were related to 1996 subordinate ratings (i.e., low scorers in 1995 tended to be low scorers in 1996, and high scorers in 1995 tended to be high scorers in 1996). A positive regression coefficient for \( Y \) would indicate that, independent of initial level of subordinate ratings, the higher an individual’s 1995 self-rating, the higher the 1996 subordinate rating tended to be. Because this effect was not expected to occur throughout the range of \( Y \) (i.e., it was only expected for overraters), the interaction between \( X \) and \( Y \) should be significant.

Regression results are shown in Table 2. Significant regression coefficients were obtained for \( X \), \( Y \), and \( X^2 \), indicating both 1995 subordinate
ratings and 1995 self-ratings contributed to the prediction of 1996 subordinate ratings, and there is some curvilinearity in the model. Contrary to expectations, however, the interaction between \( X \) and \( Y \) was not significant. The presence of higher order terms was explored, as suggested by Edwards (1994). The increment in variance explained (\( \Delta R^2 = .003 \)) was very small, so these results were not interpreted.

Figure 1 displays the model graphically. To facilitate interpretation, example data that would create a graph like this are provided in Table 3, along with predicted values based solely on regression to the mean. So as not to extrapolate the results beyond the effective range of the data, the minimum and maximum values in the graph and the table were set such that each cell in the table represented a realistic combination of self- and subordinate ratings. The graph and the table show that, at a given level of 1995 subordinate ratings, 1996 subordinate ratings increase as 1995 self-ratings increase. These results can be further illustrated by examining the slope and intercept along lines of interest in the \( X, Y \) plane. For example, the slope and intercept along the \( Y = X \) line (i.e., self-ratings are equal to subordinate ratings) can be calculated by substituting \( X \) for \( Y \) in Equation 1 (see Edwards & Parry, 1993):

\[
Z = b_0 + b_1 X + b_2 X + b_3 X^2 + b_4 X^2 + b_5 X^2 + e
= b_0 + (b_1 + b_2) X + (b_3 + b_4 + b_5) X^2 + e
\]

Thus, the slope along the \( Y = X \) line at the point \( X = 0, Y = 0 \) is given by \( b_1 + b_2 = .590 \). To see how this equation differs from that for overraters, we can substitute the expression \( X + 1 \) for \( Y \) to represent managers whose self-rating is one point higher than their subordinate rating:

\[
Z = b_0 + b_1 X + b_2 (X + 1) + b_3 X^2 + b_4 (X + 1) + b_5 (X + 1)^2 + e
= (b_0 + b_2 + b_3) + (b_1 + b_2 + b_4 + 2b_5) X + (b_3 + b_4 + b_5) X^2 + e
\]
Similarly, we can substitute the expression \( X - 1 \) for \( Y \) to represent managers whose self-rating is one point lower than their subordinate rating (underraters):

\[
Z = b_0 + b_1 X + b_2 (X - 1) + b_3 X^2 + b_4 X (X - 1) + b_5 (X - 1)^2 + e
\]

\[
= (b_0 - b_2 + b_5) + (b_1 + b_2 - b_4 - 2b_5) X + (b_3 + b_4 + b_5) X^2 + e
\]  

(5)

The slope at \( Y = X + 1 \) is given by \( b_1 + b_2 + b_4 + 2b_5 = .502 \). The slope at \( Y = X - 1 \) is given by \( b_1 + b_2 - b_4 - 2b_5 = .678 \). This indicates that 1996 subordinate scores increase as 1995 subordinate scores increase at a more rapid rate for underraters than for overraters. The other difference in the preceding three equations is in the intercept term. When \( Y = X \), the intercept is equal to \( b_0 = 0.026 \). When \( Y = X + 1 \), the intercept is equal to \( b_0 + b_1 + b_3 + b_5 = 0.081 \). When \( Y = X - 1 \), the intercept is equal to \( b_0 - b_1 + b_3 - b_5 = -0.125 \). This means that the constant in the prediction equation increases when the manager is an overrater and decreases when the manager is an underrater. In other words, at a given level of 1995 subordinate ratings, the model would predict a higher 1996 subordinate rating for an overrater than it would for an in-agreement or an underrater. Similarly, it would predict a lower 1996 subordinate rating for an underrater than it would for an overrater or an in-agreement rater.

To illustrate more concretely whether changes in subordinate ratings from year to year are positive or negative, the values in Table 3 can be compared to \( Z_{\text{reg}} \). Remember that \( Z_{\text{reg}} \) is only an approximation, so great importance should not be placed on small deviations. When \( X \) is approximately equal to \( Y \), predicted 1996 subordinate ratings are about the same as \( Z_{\text{reg}} \). The values in Table 3 show that overraters at all levels of initial performance tended to improve, in-agreement raters tended to stay the same, and underraters at moderate to high levels of initial performance tended to decline.

These results provide support for Hypothesis 1a. Predicted 1996 subordinate ratings increased as self-ratings increased at all levels of 1995 subordinate ratings. Hypothesis 1b, which predicted no gain in performance for underraters and in-agreement raters, was generally supported, although the finding that underraters showed a decline in performance from 1995 to 1996 at moderate and high levels of initial performance was unexpected.

**Change in Self-Ratings**

Hypothesis 2 predicted that the self-ratings of underraters would in-
increase and the self-ratings of overraters would decrease following feedback. This was tested by setting \( Z = 1996 \) self-ratings in Equation 1. The hypothesis suggests positive regression coefficients for \( X \) (1995 subordinate rating) and \( Y \) (1995 self-rating). A positive regression coefficient for \( Y \) would indicate that 1995 self-ratings were related to 1996 self-ratings (i.e., low self-raters in 1995 tended to be low self-raters in 1996, and high self-raters in 1995 tended to be high self-raters in 1996). A positive regression coefficient for \( X \) would indicate that, independent of initial level of self-ratings, the higher an individual's 1995 subordinate rating, the higher the 1996 self-rating tended to be. Because changes in self-ratings were expected for both overraters and underraters, there should be no interaction between \( X \) and \( Y \).

Results are shown in Table 4. Significant regression coefficients were obtained for \( X \), \( Y \), \( XY \), and \( Y^2 \). The presence of higher order terms was explored, but the increment in variance explained (\( \Delta R^2 = .003 \)) was again very small, so these results were not interpreted. The positive regression coefficients for \( X \) and \( Y \) were expected, and indicate that the self-rating would tend to increase for underraters and decrease for overraters. The positive regression coefficient for the interaction between \( X \) and \( Y \) was not expected because it indicates that the relationship between self-other agreement and change in self-ratings differs at different levels of \( X \) and \( Y \). The negative regression coefficient for \( Y^2 \) indicates a general decrease in self-ratings from year to year at all levels of \( Y \). Because negative terms become positive when they are squared, the negative coefficient decreases the predicted self-rating regardless of whether the initial self-rating is above or below the mean.

Figure 2 displays this model graphically, and example data that would create a graph like this are provided in Table 5. Table 5 also includes the predicted 1996 self-rating based solely on regression to the mean. The graph and the table show that, at a given level of 1995 self-ratings (\( Y \), 1996 self-ratings (\( Z \)) tend to increase as 1995 subordinate ratings (\( X \)) increase. In other words, managers adjusted their self-ratings according
to the feedback they received from their subordinates. This effect is not the same throughout the range of $Y$, however. $Z$ tends to rise faster as $Y$ increases.

The effects of $X$ at different levels of $Y$ can also be examined by looking at the slope of the line predicting $Z$ at different levels of $Y$. The slope of this line is defined as $b_1 + b_4 Y$. Thus, at $Y = -0.80$, the slope is $-0.024$. The slope becomes less negative as $Y$ increases, and slope $= 0$ at $Y = -0.64$. The slope is $0.100$ at $Y = 0$, and $0.224$ at $Y = 0.80$.

When comparing predicted 1996 self-ratings with $Z_{reg}$, it can be seen that there was a general decline in self-ratings, so both overraters and in-agreement raters tended to decrease their self-ratings from 1995 to 1996. At $Y = .60$ and below, self-ratings tended to exceed $Z_{reg}$ at all levels of $X$, but the amount by which subordinate ratings exceed self-ratings has no effect on predicted self-ratings. These results provide general support for Hypothesis 2, with some qualifications. Managers did tend to adjust their self-ratings from year to year according to the level of their subordinate ratings, but this effect was strongest at high levels of 1995 self-ratings, and disappeared at low levels of 1995 self-ratings.

Interrater Agreement

Two operationalizations of interrater agreement were used, because it was unclear which would better represent the concept of interrater agreement as perceived by the feedback recipient. They were (a) mean $r_{agg}$ across rating instrument items, and (b) mean variance across items. Higher $r_{agg}$ and lower variance indicate greater interrater agreement, so Hypothesis 3 predicted a positive relationship between mean $r_{agg}$ and change in subordinate ratings and a negative relationship between mean variance and change in subordinate ratings. The moderating effect of interrater agreement was tested with the following model (cf. Edwards, 1994):

$$ Z = b_0 + b_1 X + b_2 Y + b_3 X^2 + b_4 XY + b_5 Y^2 + b_6 A $$

$$ + b_7 AX + b_8 AY + b_9 A X^2 + b_{10} AXY + b_{11} AY^2 + e $$

(6)

where $A =$ interrater agreement and the other terms are defined as in Equation 1. If the test for the addition of interrater agreement to the model is significant, agreement contributes to the prediction of 1996 subordinate ratings beyond what is predicted by 1995 subordinate and self-ratings. If the test for the last five terms as a set is significant, a moderating effect for interrater agreement is suggested (Edwards, 1994). The regression coefficients for this model are given in Table 6. The increase in $R^2$ by adding interrater agreement alone was significant both when
TABLE 6
Polynomial Regression Results When Testing the Moderating Effect of Interrater Agreement on the Relationship Between Self-Other Agreement and Change in Subordinate Ratings

<table>
<thead>
<tr>
<th>Variable</th>
<th>( B )</th>
<th>SE</th>
<th>t</th>
<th>( B )</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>.082</td>
<td>.034</td>
<td>2.41*</td>
<td>-.026</td>
<td>.026</td>
<td>0.99</td>
</tr>
<tr>
<td>X</td>
<td>.531</td>
<td>.093</td>
<td>5.70***</td>
<td>.548</td>
<td>.052</td>
<td>10.33***</td>
</tr>
<tr>
<td>Y</td>
<td>.160</td>
<td>.067</td>
<td>2.40*</td>
<td>.061</td>
<td>.048</td>
<td>1.29</td>
</tr>
<tr>
<td>( X^2 )</td>
<td>.279</td>
<td>.115</td>
<td>2.43*</td>
<td>-.007</td>
<td>.089</td>
<td>-0.08</td>
</tr>
<tr>
<td>( XY )</td>
<td>.110</td>
<td>.135</td>
<td>0.81</td>
<td>-.060</td>
<td>.103</td>
<td>-0.58</td>
</tr>
<tr>
<td>( Y^2 )</td>
<td>-.125</td>
<td>.083</td>
<td>-1.51</td>
<td>-.008</td>
<td>.067</td>
<td>-0.12</td>
</tr>
</tbody>
</table>

\[ R^2 = .280 \]

\[ R^2 = \Delta R^2 = .003 \]

Note: \( X = 1995 \) subordinate rating. \( Y = 1995 \) self-rating. \( A = \) Interrater agreement. \( N = 1,888 \). All regression coefficients are for the full model.

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

A was operationalized as average \( r_{wg} \) (\( \Delta R^2 = .003, F = 6.87, df = 1, 1,881, p < .01 \)) and when A was operationalized as average variance (\( \Delta R^2 = .005, F = 12.10, df = 1, 1,881, p < .001 \)). The subsequent increase in \( R^2 \) by adding the last five terms was also significant both when A was operationalized as average \( r_{wg} \) (\( \Delta R^2 = .003, F = 6.80, df = 5, 1,876, p < .001 \)) and when A was operationalized as average variance (\( \Delta R^2 = .002, F = 5.55, df = 5, 1,876, p < .001 \)). Because the effect sizes were so small, it is probably best to conclude that interrater agreement had a small effect on 1996 subordinate ratings, but it did not have a meaningful moderating effect.

Examination of the regression coefficients in Table 6 shows that the sign of the regression coefficient for the measure of interrater agreement was opposite the hypothesized direction. The positive regression coefficient associated with average variance indicates that the more variance there was in the ratings (i.e., the less interrater agreement), the higher the 1996 subordinate ratings tended to be. These results therefore fail to support Hypothesis 3.

TABLE 7
Polynomial Regression Results When Testing the Moderating Effect of Participation in 1994

<table>
<thead>
<tr>
<th>Variable</th>
<th>( B )</th>
<th>SE</th>
<th>t</th>
<th>( B )</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-.014</td>
<td>.031</td>
<td>-.46</td>
<td>-.006</td>
<td>.032</td>
<td>-0.18</td>
</tr>
<tr>
<td>X</td>
<td>.570</td>
<td>.058</td>
<td>9.76***</td>
<td>.339</td>
<td>.060</td>
<td>2.32*</td>
</tr>
<tr>
<td>Y</td>
<td>.093</td>
<td>.048</td>
<td>1.97*</td>
<td>.435</td>
<td>.052</td>
<td>8.36***</td>
</tr>
<tr>
<td>( X^2 )</td>
<td>.167</td>
<td>.107</td>
<td>1.57</td>
<td>-.101</td>
<td>.104</td>
<td>-0.97</td>
</tr>
<tr>
<td>( XY )</td>
<td>.119</td>
<td>.127</td>
<td>0.93</td>
<td>.378</td>
<td>.129</td>
<td>2.93**</td>
</tr>
<tr>
<td>( Y^2 )</td>
<td>-.211</td>
<td>.080</td>
<td>-2.65**</td>
<td>-.132</td>
<td>.089</td>
<td>-1.49</td>
</tr>
</tbody>
</table>

\[ R^2 = .280 \]

\[ R^2 = \Delta R^2 = .004*** \]


All regression coefficients are for the full model.

Additional Moderators

Additional moderators were tested using the same procedure used for testing interrater agreement as a moderator. Because 1995 was the second administration of the upward feedback program, a possible moderating variable was whether or not the manager participated in the program during the first administration in 1994. A dummy variable was created that was coded 1 = participated in 1994 and 0 = did not participate in 1994. Equation 5 was then tested, with A = participation in 1994. Regression coefficients are shown in Table 7. The change in \( R^2 \) when adding the participation variable was small (\( \Delta R^2 = .004, F = 9.77, df = 1, 1,881, p < .01 \)), as was the change in \( R^2 \) when adding the interaction terms (\( \Delta R^2 = .003, F = 7.66, df = 5, 1,876, p < .001 \)). The positive regression coefficient for the participation variable indicates that there was a slight tendency for managers who participated in 1994 to improve more from 1995 to 1996 than did managers who did not participate in 1994. The significant interaction indicates that this effect was not constant at all levels of 1995 subordinate ratings. Again, this was a very small effect.
Table 7 also contains the regression coefficients for the test of the moderating effect of participation in 1994 on the relationship between self-other agreement and 1996 self-ratings. The change in $R^2$ when adding the participation variable was again very small ($\Delta R^2 = .002, F = 4.62, df = 1, 1,825, p < .05$), as was the change in $R^2$ when adding the interaction terms ($\Delta R^2 = .003, F = 8.75, df = 5, 1,820, p < .001$).

Another potential moderating variable was comparison to the norm group. In the feedback report, each manager was provided with the average score received by everyone at his or her level. It is possible that standing relative to the norm group could influence future subordinate or self-ratings. A dummy variable was created coded 1 = above the norm group and 0 = below the norm group. No effect was found for norm group comparison when the dependent variable was 1996 subordinate ratings or when the dependent variable was 1996 self-ratings.

**Discussion**

**Change in Subordinate and Self-Ratings**

All hypotheses about the relationship between self-other agreement and changes in subordinate and self-ratings following feedback were generally supported. As suggested by Korman's (1970, 1976) self-consistency theory, subordinate ratings of overraters tended to increase from 1995 to 1996, and self-ratings of overraters tended to decrease from 1995 to 1996. Managers who rated themselves at about the same level as their subordinates rated them tended to have a slight increase in subordinate ratings and a slight decrease in self-ratings from 1995 to 1996. Managers who underrated themselves in 1995 tended to have lower average subordinate ratings and higher self-ratings in 1996.

The finding that underraters tended to decrease their performance from one year to the next was not predicted by self-consistency theory. It was expected that these managers would be more likely to adjust their self-perceptions to be in line with others' perceptions than to adjust their behavior to bring others' perceptions in line with their own. Although subordinate ratings of underraters tended to decrease, it is unlikely that these managers decreased their performance for the purpose of being consistent with others' perceptions. A possible explanation is that the performance of underraters declined because they had little motivation to even maintain their current level of performance. Ashford (1989) suggested that individuals evaluate their behavior relative to their self-ratings. If underraters tend to have low aspiration levels (Atwater & Yammarino, 1997), feedback indicating they are exceeding their own standards may have a demotivating effect and lead them to decrease their efforts. Self-consistency theory probably accounts for why most underraters raised their self-ratings, but is not a likely explanation for why their subordinate ratings decreased.

Another unexpected finding was that the rate of change in self-ratings was not constant throughout the range of initial self-ratings. Initially low self-raters increased their self-ratings beyond what would be expected by regression to the mean, but the amount by which they underrated themselves had no effect on the increase in self-ratings. The amount by which self-ratings differed from subordinate ratings really only mattered at moderate to high levels of initial self-ratings. It may be that underraters were aware that their performance had declined and took that into account when making their 1996 self-ratings. On the one hand, underraters were expected to increase their self-ratings by a greater amount the more they underrated themselves. On the other hand, subordinate ratings decreased to a greater extent the more managers underrated themselves. The expected increase in self-ratings associated with subordinate ratings exceeding self-ratings may have been offset by the systematic decrease in performance associated with subordinate ratings exceeding self-ratings, if managers used both kinds of information when arriving at their 1996 self-ratings. A similar effect, however, was not found for overraters. As the extent of overrating increased, subordinate ratings continued to increase and self-ratings continued to decrease.

**Alternative Explanations for Changes in Subordinate Ratings**

There are other theories besides self-consistency theory that could explain the results of this study. For example, goal-setting theory (Locke & Latham, 1990) and control theory (Carver & Scheier, 1981, 1982) would make similar predictions about the change in performance as self-consistency theory, if it is assumed that self-ratings represent the standard against which managers compare themselves (cf. Ashford, 1989). Overraters have a discrepancy between their feedback and their goal (i.e., their self-ratings), so their improvement could be explained as either (a) striving to achieve the goal (goal-setting theory), or (b) striving to reduce the discrepancy between current performance and the goal (control theory). In-agreement raters and underraters have achieved or surpassed their goal, so they would not be expected to improve. Equating self-ratings and goals, however, is a very large assumption.

**Alternative Explanations for Changes in Self-Ratings**

There are alternative explanations for why overraters tended to de-
cease their self-ratings and underraters tended to increase their self-ratings from 1995 to 1996. It is possible that managers simply wanted to appear to be sensitive to feedback. All managers were required to share their feedback report with their immediate supervisor, so they may have consciously tried to change their self-ratings in 1996 to appear to be responsive to the feedback they received in 1995. This explanation cannot be ruled out.

Another possibility is that the decrease in self-ratings for overraters and the increase in self-ratings for underraters had less to do with self-other rating discrepancies than with the normative information provided with the 1995 feedback (i.e., the mean rating for all participants at the level of the ratee). According to social comparison theory (Festinger, 1954), it is easier to judge one’s self on some characteristic by comparing oneself to others than to judge oneself with no information about others. Indeed, people do make more accurate self-assessments when social comparison information is available. For example, self-ratings of performance are more similar to others’ ratings when average ratings are known to the self-raters than when they are not (Farh & Dobbins, 1989; Heneman, 1989; Mabe & West, 1984). Therefore, it is possible that the norms were informative to managers who had rated themselves especially high or especially low in 1995, and these managers adjusted their self-ratings to be closer to average in 1996. We found no moderating effect of being above or below the norm in our data, however.

It is also possible that differences in self-ratings reflect actual changes in performance. In other words, decreased self-ratings may reflect actual declines in performance among overraters and increased self-ratings may reflect performance improvement among underraters. This explanation is very unlikely, given the fact that subordinate ratings tended to increase for overraters and decrease for most underraters from 1995 to 1996. This is a possible explanation, however, for the relationship between the change in self-ratings and the change in subordinate ratings among managers with initially low self-ratings.

**Interrater Agreement**

The hypothesis regarding the effect of interrater agreement on performance improvement was not supported. There was a small negative relationship between the extent to which raters agreed on their ratings of a manager and the extent to which that manager improved following feedback. A possible explanation is that less agreement among raters increases the salience of the feedback. More salient feedback may increase the attention paid to self-other discrepancies (London & Smither,

1995). Another possibility is that managers may feel helpless and unmotivated to improve if all their subordinates rate them poorly, but may feel hopeful and more motivated if at least some of them give relatively high ratings.

There is another likely explanation for this result. The overall level of performance in these data was quite high (mean 1995 subordinate rating = 3.98 on a 5-point scale), so ratings of 1 and 2 were unusual. This is consistent with the negative skew typically found in research on performance ratings (Landy & Farr, 1980). A very low rating is an outlier, and this outlier would have two effects: (a) decrease the mean, and (b) increase the variance. Lower overall ratings would tend to have greater variance, which is shown by the correlation of .55 between 1995 subordinate rating and average variance. Greater variance in the ratings means there is more error, so more regression to the mean would be expected. For example, consider two managers who were each rated by six subordinates in 1995, each with a mean rating of 2.50. If Manager A received three ratings of 2 and three ratings of 3, and Manager B received three ratings of 1 and three ratings of 4, there would be more variance in the ratings of Manager B. The three ratings of 1 are less likely to be repeated in 1996 than are the three ratings of 4, because they are outliers. The fact that the raters in 1996 are not necessarily all the same raters as in 1995 makes it even less likely that a rating of 1 would be repeated. Thus, one would expect more regression to the mean for Manager B than for Manager A, and the mean subordinate rating in 1996 would likely be higher for Manager B than for Manager A. This makes it appear that greater variance in the ratings leads to greater performance improvement, but it really only reflects the differential effects of regression to the mean based on the amount of error in the ratings. A great deal more research on this issue is necessary before definitive conclusions can be drawn about the value of presenting rates with the variance or range of ratings, but less agreement probably does not lead to greater improvement.

**Participation at Time 1**

The year for which the first round of data was available for this study was the second year of administration. Whether the manager participated in the program during the first administration was found to be related to both change in subordinate ratings and change in self-ratings. Managers who participated in the first administration tended to improve more from the second to the third administration than did managers who were participating for the first time. This suggests that feedback is at least as motivating the second time around as the first.
Implications

This study has clear implications for the motivating effect of self-ratings. The results suggest that managers who receive upward feedback will tend to improve their performance to a greater extent the more their self-ratings exceed their subordinate ratings. If self-ratings are lower than subordinate ratings, managers tend to decrease their performance as the discrepancy increases. Upward feedback therefore does not necessarily have positive effects on performance for all managers. This study identified specific conditions under which upward feedback would not be expected to lead to improvements in performance. These conditions were consistent with the results of Smith et al. (1995) and Atwater et al. (1995), but this study had a much larger sample and used polynomial regression rather than difference scores to test relationships. This study was the first to show that overratings at all levels of initial performance tend to improve, not just overratings at a low initial level of performance.

Previous studies have also not found the decrement in performance associated with underraters that was found in this study. Atwater et al. (1995) found no significant difference in follower ratings before and after feedback. This result, however, was based on a trichotomization of self-other agreement that confounded agreement and initial level of performance. Their analyses do not tell us whether underraters would be expected to have a more negative change in performance than in-agreement raters who are at the same level of initial performance. Their reported results, however, do suggest that this is the case. Although no significant change in follower ratings was found for either under- or in-agreement raters, the follower ratings of underraters declined from a mean of 4.15 to a mean of 4.10, while the follower ratings of in-agreement raters increased from a mean of 3.90 to a mean of 4.05. The difference between underraters and in-agreement raters was significant before feedback and not significant following feedback. It is difficult to disentangle the effects of initial level of performance and regression to the mean on these results, but it appears that Atwater et al.'s results are not necessarily inconsistent with ours. More research is needed to understand why upward feedback tends to be less beneficial for underraters.

Previous studies have emphasized the initial level of performance as the key determinant of performance improvement, citing goal-setting and control theories as the proper explanatory framework (e.g., Reilly et al., 1996; Smith et al., 1995; Walker, 1997). This study found that self-other agreement affected subsequent performance improvement at all initial levels of performance, suggesting that self-consistency theory also provides a useful perspective when doing research on upward feedback programs that incorporate self-ratings. Future research is necessary to estimate the relative contributions different kinds of goals and the desire for self-consistency make to motivation to improve. Establishing boundary conditions under which hypotheses made by different theories hold is one way of increasing our understanding of the processes by which upward feedback operates.

The results of this study make it appear that upward feedback has very little effect on overall performance improvement across raters. The standardized difference between overall mean subordinate ratings in 1996 and 1995 (Cohen's $d$) was only 0.05. Gains in performance made by overraters were offset by declines among underraters. At first glance, this appears inconsistent with other research in this area. For example, Atwater et al. (1995) found an increase in mean ratings from 3.77 prefeedback to 3.99 postfeedback. Other studies, however, have found small or negative differences between consecutive administrations. Smith et al. (1995) reported a $d$ of 0.16 between Time 1 and Time 2 ratings, but mean scores declined overall between Time 2 and Time 3 ($d = -0.07$) before rising again at Time 4 ($d = 0.19$; Reilly et al., 1996). Walker (1997) found the following effect sizes between consecutive administrations: 0.00 between Time 1 and Time 2, 0.22 between Time 2 and Time 3, 0.09 between Time 3 and Time 4, and 0.02 between Time 4 and Time 5. In our study, there was a decline in performance for managers who did not participate in the first round of feedback (i.e., between Time 1 and Time 2), $d = -0.08$, $t$ (274) = -1.27, ns. There was a small increase in performance for managers who did participate at Time 1 (i.e., between Time 2 and Time 3), $d = 0.07$, $t$ (1,883) = 3.01, $p < .01$.

The results of this study suggest an explanation for the small overall increase in performance compared to other studies. Self-ratings typically exceed ratings made by others (Atwater et al., 1995; Smith et al., 1995; see also Harris & Schaubroeck, 1988), but the mean self-rating was lower than the mean subordinate rating in our study. Performance tends to improve primarily for overraters, so it is not surprising that our overall level of improvement is smaller than in previous studies, because the proportion of overraters in this study was smaller than in previous studies. As with other feedback interventions (see Kluger & DeNisi, 1996), we cannot assume that feedback will lead to positive changes in performance for everyone. Instead we must examine the conditions under which upward feedback is beneficial. The present study identified one such condition: Upward feedback is beneficial when the ratee learns that his or her self-ratings exceed the ratings made by his or her subordinates.
The results of this study suggest the value of examining the method by which feedback is presented to managers. Many feedback instruments highlight the discrepancy between self-ratings and ratings made by others (Van Velsor & Leslie, 1993). It may be preferable, however, to tailor the feedback to the individual depending on the level of self–other agreement. Emphasis on self–other discrepancies may be appropriate for overraters, but it could do nothing or decrease the motivation of in-agreement raters and underraters. For in-agreement raters and underraters, the self–other discrepancy should perhaps be de-emphasized, with more emphasis placed on the absolute level of the ratings or comparison to a norm group. These raters must focus on a different standard of comparison than their self-ratings, such as how they are doing relative to optimal performance or to other managers in the organization. Research is needed on the extent to which upward feedback recipients pay attention to different standards of comparison (e.g., self-ratings, norm groups, previous performance) and the conditions under which the salience of different standards can be manipulated.

**Limitations**

The biggest limitation of this study was that we could not determine what kinds of improvement goals managers set after receiving feedback. The results were consistent with the predictions of self-consistency theory, but we do not know for sure that overraters were motivated to bring others’ perceptions in line with their own. Because of the archival nature of the data, we were unable to ask managers what they paid attention to during feedback. This made it impossible to evaluate the relative explanatory power of different theories for our results. The fact that most research on multiple perspective feedback must be conducted on archival data is a serious problem that must be addressed in future research in this area (J. W. Smither, personal communication, April 11, 1997).

Another potential limitation is that the first wave of feedback for which data were available (1995) was the second wave of feedback for most of the managers in the sample. We found that participation at Time 1 had a positive effect or no effect on performance improvement, depending on the level of performance at Time 2. It is not known, however, how the motivating effect of self-ratings changes over multiple administrations. For example, if overraters are highly motivated to improve after the first or second round of feedback, would they continue to be motivated if they are still overraters after three, four, or five rounds of feedback? To examine these questions, this study should be replicated across several administrations.

Another limitation of this study is the lack of a no-feedback control group. It is difficult to draw conclusions about changes in self- or subordinate ratings as a result of feedback because we do not know how these ratings would have changed in the absence of feedback. Smither et al. (1995) found that managers who improved did so regardless of whether or not they received upward feedback, and Reilly et al. (1996) found that performance improvement was not related to the number of times participants received feedback. The focus of this study, however, was not on the amount of improvement following feedback, but on the differences in the amount of improvement for different types of managers (e.g., overraters, underraters). We can conclude that ratings changed in a positive direction for overraters and that ratings changed in a negative direction for underraters, but we cannot conclude anything about the absolute level of the change because of the lack of a control group. For example, it is possible that the performance of managers who did not receive feedback declined from 1995 to 1996 by a greater amount than did the performance of underraters. If this were the case, the feedback would have had a positive effect on underraters, even though their ratings decreased. We are not aware of anything occurring in the organization that would have been likely to influence the performance of the managers independent of the feedback program (e.g., organizational development efforts, training programs), so it is unlikely that any systematic change in performance occurred outside of the feedback program. Nevertheless, the absolute level of effect sizes should be interpreted cautiously, and it is the differences between effect sizes for different types of managers that should be of primary interest.

Related to this is the question of regression to the mean. In the absence of good information regarding the level of test-retest reliability for self- and subordinate ratings, it was necessary to base our predicted values based solely on regression to the mean on reasonable estimates. This also calls into question the absolute level of the effect sizes, but has no effect on the differences between effect sizes. In future longitudinal studies, researchers should make every attempt to evaluate the test-retest reliability of their measures so the effect of regression to the mean can be estimated with more confidence.

**REFERENCES**


