Raters Who Pursue Different Goals Give Different Ratings

Kevin R. Murphy, Jeanette N. Cleveland, Amie L. Skattebo, and Ted B. Kinney
Pennsylvania State University

J. N. Cleveland and K. R. Murphy (1992) suggested that phenomena such as rater errors and interrater disagreements could be understood in terms of differences in the goals pursued by various raters. We measured 19 rating goals of students at the beginning of a semester, grouped them into scales, and correlated these with teacher evaluations collected at the end of the semester. We found significant multiple correlations, both within classes and in an analysis of the pooled sample (adjusting for instructor mean differences, incremental $R^2 = .08$). Measures of rating goals obtained after raters had observed a significant proportion of ratee performance accounted for variance (incremental $R^2 = .07$) not accounted for by measures of goals obtained at the beginning of the semester.

The reliability and validity of performance ratings have long been a source of concern to researchers and practitioners alike. Variables ranging from rating scale formats to the cognitive limitations of raters have been identified as potential explanations for the apparently low quality of rating data (Bretz, Milkovich, & Read, 1992; Landy & Farr, 1980; Saal, Downey, & Lahey, 1980). The performance appraisal literature includes a number of suggestions for improving performance appraisals (e.g., rater training, scale refinement), most of which are based on the assumption that raters do not have the skills, rating scales, or information needed to accurately evaluate performance (see Landy & Farr, 1980, and Murphy & Cleveland, 1991, 1995, for reviews of this research). It is possible, however, that “rater errors” and other psychometric deficiencies in rating might not be the result of errors or limitations in the rater’s capacity, but rather might reflect the effects of strategic decisions on the part of raters about the sorts of ratings they should record. Cleveland and Murphy (1992) suggested that raters pursue a number of goals when completing performance appraisals (e.g., using ratings to maintain harmony within the workgroup, to motivate subordinates to perform better in the future, etc.), and these goals may lead them to give ratings that can appear psychometrically suspect. For example, a rater who is lenient (i.e., who gives high ratings) might not be making a judgment error at all but rather might be making a carefully calculated decision that it is better to give all subordinates high ratings than to give low ratings to poor-performing subordinates. Uniformly high ratings might lead to better pay for subordinates, more harmony in the workgroup, better supervisor–subordinate relations, and so forth, whereas more accurate ratings may lead to resentment, low motivation, and friction in the group.

Murphy and Cleveland’s (1995) summary of performance appraisal research and theory focused extensively on the role of rater goals. They noted that there is considerable anecdotal support for the hypothesis that the goals pursued by raters influence the ratings they give, in areas ranging from studies of performance rating in the military (Bjerke, Cleveland, Morrison, & Wilson, 1987) to descriptions of political manipulation of performance appraisals of managers and executives (Gioia & Longenecker, 1994; Longenecker, Sims, & Gioia, 1987). Furthermore, these studies suggest that raters are able to articulate the goals that are most important to them when rating their subordinates, and that their descriptions of the goals they intend to pursue predict their rating behavior (e.g., the likelihood of giving high ratings). Finally, as Harter (1986) and others (e.g., Bjerke et al., 1987) have noted, the goals raters actually pursue are not always the same as the goals the organization would like them to pursue, and conflicts between the official purpose of performance appraisal systems and the ways raters actually use these systems can substantially affect the utility of performance appraisal.

Rater goals are thought to play an important part in understanding the rater’s motivation when completing performance appraisals. Rating goals reflect end states a rater aims toward when completing performance appraisals (cf. Bandura & Walters, 1963; Beach, 1990; Beach & Mitchell, 1987; Pervin, 1983). That is, when raters complete performance appraisals, it is assumed that they have specific (and possibly multiple) goals in mind and that they intend to provide ratings that are consistent with these goals. Thus, a rater who intends to use performance appraisal as a means of motivating his or her subordinates will give ratings that are most likely to encourage future performance, not necessarily ratings that accurately reflect past performance. Many of the support systems and interventions in performance appraisal (e.g., training, scale refinement) appear to be based on the questionable assumption that raters are trying their best to provide accurate ratings but that they lack the skills and tools to do the job. It is more likely that raters are capable of giving accurate ratings but are often motivated to give ratings that advance goals such as motivating their subordinates or preserving group harmony (Banks & Murphy, 1985).

It is unlikely that goals are the sole determinant of ratings; there is evidence that ratings are influenced by many aspects of the rating process, by the personalities and the cognitive processes of raters and, of course, by the behavior of ratees (Landy & Farr, 1980). It is also likely that rater goals are at least partially con-
founded with other rater attributes. For example, some raters might be consistently lenient or severe, not because of their assessments of the costs and benefits of giving high or low ratings but rather because of their personalities and temperaments. Some raters might believe that rigorous evaluations challenge and motivate their subordinates, whereas others might believe that giving high ratings shows support and empathy. Regardless of the factors that lead raters to adopt different rating goals, the research reviewed above suggests that these goals will influence the ratings they give.

To date, there have been few direct empirical studies of the links between rater goals and performance ratings. The purpose of the present study is to provide an empirical test of the proposition that the goals raters claim to emphasize when evaluating performance are related to the ratings they give.

**The Present Study**

The present study was conducted in the context of university teacher evaluation, which provides a conservative test of goal-related hypotheses. Unlike traditional performance appraisals, teacher ratings do not have immediate consequences for the rater; teacher ratings are anonymous, and raters have little to gain by giving overly positive ratings and little to lose by giving overly harsh ones. This lack of immediate consequences is likely to produce weaker goal effects than in the context of supervisory ratings, in which raters are quite likely to suffer negative consequences if they give low ratings and achieve valued rewards if they give higher ratings (Cleveland & Murphy, 1992). Thus, a demonstration of goal-rating relationships in the context of university teacher ratings might argue for even stronger links in contexts in which raters have stronger incentives to give particular ratings.

One advantage of testing goal hypotheses in the context of teacher rating is that the design of teacher evaluation systems allows us to at least partially control the effects of ratee performance on raters’ goals and on the relationship between goals and ratings. In traditional supervisory performance appraisals, it is possible that the goals adopted by raters and the relationship between goals and ratings are both influenced by the performance level of the ratee (Murphy & Cleveland, 1995). In settings where each rater is evaluated by a different rater, it is impossible to separate the roles of ratee performance versus rater goals in determining performance ratings. In our study, multiple raters in each class observed and evaluated essentially the same sample of performance when evaluating their instructor. This allows us to substantially control the effects of ratee behavior on raters’ goals, providing an unambiguous test of the hypothesis that raters who observe the same performance, but who pursue different goals, when completing performance appraisals will tend to give different ratings.

In this study, our first hypothesis was as follows:

**Hypothesis 1:** Measures of the rating goals most strongly emphasized by raters will account for a substantial portion of the variance in the performance ratings they assign.

Our design allowed us to obtain information about some raters’ goals both before they had a meaningful chance to observe the ratee’s performance (at the beginning of the semester) and after they had observed a substantial sample of the ratee’s performance (at the end of the semester). On the basis of Murphy and Cleveland’s (1995) prediction that rater goals would be influenced by the ratee’s level of performance, we hypothesized that the importance of rating goals would change over time and that these changes would be related to the rater’s final evaluation of the ratee.

In particular, we hypothesized the following:

**Hypothesis 2:** Measures of goal importance obtained after the rater has observed the ratee will account for variance in performance ratings not accounted for by ratings of goal importance collected before observing the ratee’s performance.

**Method**

We first conducted a pilot study to provide data for scale development and to provide an initial estimate of the strength of goal-rating relationships. We then carried out our main study in which measures of goal importance were obtained before the raters had an opportunity to observe a substantial sample of ratee performance; goal importance ratings were used to predict evaluations of the instructors’ performance collected several months later.

**Pilot Study**

Students in seven separate classes (N ranged from 19 to 187), each with a different instructor, were asked to complete a rating goal questionnaire at the same time they completed a standardized university-wide, end-of-semester teacher evaluation form.

**Procedures and participants.** Over 90% of respondents were undergraduates in their first 3 years of college. Questionnaires were distributed in counter-balanced order, with half of the participants completing the goal questionnaire first and half completing the teacher rating form first.

**Goal questionnaire.** A 19-item questionnaire was developed to assess the goals pursued by raters in evaluating their instructors; participants were asked to indicate the extent to which they agreed with the statement that each of the 19 goals influenced or was important in the evaluation of their instructor’s performance by using a 5-point Likert-type scale, ranging from 1 (strongly disagree) to 5 (strongly agree). In developing our questionnaire, we used items and item statistics from a study by Kerst (1993), which investigated performance rating as a function of rating purpose, ratee performance level, and rater goals. As part of her study, Kerst (1993) asked managers and supervisors in a medium-sized organization to generate examples of goals they might pursue when completing performance evaluations. She then had other supervisors retranslate the goals into categories based on the purpose for which appraisals may be conducted (to provide feedback or to make administrative decisions) and the performance level of the ratee (poor, average, outstanding).

The 19 goal items presented in Table 1 could be grouped on a conceptual basis around four main themes. Eight of the items described goals that related to identifying the instructor’s weaknesses (Cronbach’s α = .91). Six items described goals that related to identifying and conveying information about strengths (Cronbach’s alpha = .74). Three items dealt with providing fair ratings that were clearly tied to performance standards (Cronbach’s α = .60), and two items dealt with motivating raters (Cronbach’s α = .67). A confirmatory factor analysis of goal ratings obtained in this study showed that this model fits the goal-rating data very well (normed fit index [NFI] = .97, comparative fit index [CFI] = .98, root-mean-square error of approximation [RMSEA] = .09); we created scales to represent each of these correlated factors. Intercorrelations among the four factors ranged from .26 to .53, with a median value of .46.

**Teacher rating forms.** The teacher evaluation form used by the university included three items that referred specifically to the effectiveness of
the instructor and/or the course (“Overall, I would rate this course as good”; “The teacher organized the course effectively”; “Overall, I would rate this teacher as good”). In addition to these three items, we asked students to complete a one-item overall rating of their instructor’s performance in the course (i.e., “The instructor performed very well in this course”). Raters responded using the same 5-point scale as was used in the goal questionnaire. Ratings on all four items were highly correlated, and a four-item composite provided a highly reliable index of instructor performance (Cronbach’s $\alpha = .94$).

Links between goals and ratings—pilot study. We first examined relationships between goals and ratings within each class. Scores on the four goal scales were significantly related to ratings in all of the classes, with $R^2$ values ranging from .07 to .36 and a median value of .16. Relationships between goal ratings and teacher evaluations were not influenced by the order in which questionnaires were completed; $R^2$ values did not vary by more than .02 across the two order conditions.

We next examined relationships between goals and performance ratings in the total sample (i.e., $N = 895$) by using hierarchical regression. We first entered a set of dummy-coded variables representing the seven instructors and found that differences in the mean ratings received by the seven instructors accounted for 18.7% of the variance in teacher ratings. We then entered scores on the four goal scales and found a significant and substantial proportion of variance explained (incremental $R^2 = .24$) over and above the variance explained by instructor effects. These analyses suggest that information about rater goals can be used to predict ratings.

Main Study

Our pilot study suggests that rater goals are indeed related to the ratings raters give. However, because goal importance ratings are collected at the same time as performance ratings, it is possible that these links are spurious, either due to priming effects (e.g., the evaluations one has already formed of the instructor might influence ratings of goal importance) or common method effects (i.e., ratings of goal importance and of teacher performance are collected at the same time). To reduce the chance that ratings of goal importance are substantially influenced by participants’ evaluations of the teacher’s performance or that completing the rater goal questionnaire affects the rater’s evaluations of their instructor’s performance, we obtained ratings of goal importance at the very beginning of the semester, before students had an opportunity to observe the performance of the instructor they would evaluate, and obtained teacher ratings several months later, at the end of the semester. We also obtained a second set of ratings of goal importance at the end of the semester (after viewing the instructor’s performance) from a subsample of participants.

Ratings of goal importance were collected during the first week of the semester ($N = 303$) and ratings of teacher performance ($N = 232$) were collected at the end of the semester (14 weeks later) in five separate courses, using the same goal questionnaires and teacher rating forms as were used in the pilot study. Virtually all of our participants were undergraduates in their first 3 years of college.

We were able to match goal importance and teacher ratings for 186 participants. A second set of ratings of goal importance was also collected from a subsample of students at the end of the semester ($N = 154$; $N = 125$ matched with goal ratings collected at Time 1).

Results

All five instructors received high ratings, with a mean rating of 4.15 on our four-item composite scale and individual instructor ratings means ranging from 4.03 to 4.22. Differences across instructors accounted for 14% of the variance in teacher ratings. Mean, standard deviations, and correlations among goal impor-

<table>
<thead>
<tr>
<th>Goal</th>
<th>Weakness</th>
<th>Strength</th>
<th>Fairness</th>
<th>Motivating</th>
<th>$R^2$–Time 1</th>
<th>$R^2$–Total*</th>
<th>$\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Identify areas in which the instructor might need improvement</td>
<td>1.00</td>
<td>1.00</td>
<td>.01*</td>
<td>.11*</td>
<td>.10*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Rate my instructor fairly</td>
<td>1.26</td>
<td>1.00</td>
<td>.01*</td>
<td>.11*</td>
<td>.08*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Identify areas where the instructor needs more training</td>
<td>1.40</td>
<td></td>
<td>.03*</td>
<td>.13*</td>
<td>.10*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Convey my satisfaction with the instructor’s performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Identify areas that the instructor should focus on improving</td>
<td>1.38</td>
<td></td>
<td>.001</td>
<td>.12*</td>
<td>.12*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Indicate where the instructor fell short in terms of performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Give my instructor a rating that she or he will realize is based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Identify my instructor’s strengths and weaknesses</td>
<td>1.45</td>
<td></td>
<td>.008</td>
<td>.09*</td>
<td>.09*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Highlight my instructor’s performance so that his or her success</td>
<td>1.19</td>
<td></td>
<td>.001</td>
<td>.01</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Improve my instructor’s confidence</td>
<td>1.00</td>
<td></td>
<td>.001</td>
<td>.01</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Make it clear to my instructor that there is room for</td>
<td>1.33</td>
<td></td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Identify my instructor’s performance deficiencies</td>
<td>1.68</td>
<td></td>
<td>.02</td>
<td>.04*</td>
<td>.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Challenge my instructor to improve his or her performance</td>
<td>1.49</td>
<td></td>
<td>.003</td>
<td>.01</td>
<td>.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Clarify expected performance levels to the instructor</td>
<td>.45</td>
<td></td>
<td>.000</td>
<td>.00</td>
<td>.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Evaluate the instructor in a manner that clearly indicates what</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Indicate where the instructor has exceeded performance expectations</td>
<td>1.15</td>
<td>.01</td>
<td>.11*</td>
<td>.10*</td>
<td>.10*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17. Encourage the instructor’s current level of performance</td>
<td>1.10</td>
<td></td>
<td>.00</td>
<td>.08*</td>
<td>.08*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. Encourage the instructor to improve performance</td>
<td>1.19</td>
<td></td>
<td>.01</td>
<td>.00</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19. Motivate the instructor</td>
<td>1.50</td>
<td>.09*</td>
<td>.10*</td>
<td>.01</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* $R^2$ using both Time 1 and Time 2 ratings as predictors, $N = 125$.
* $p < .05$. 

Table 1
Relationship Between Time 1 Goal Importance, Change in Goal Importance, and Performance Ratings
tance ratings collected at the beginning and the end of the semester and composite performance ratings are shown in Table 2.

**Relationships Between Goals and Ratings**

We first examined the relationship between ratings of goal importance collected at the beginning of the semester and ratings of teacher performance obtained at the end of the semester separately in each class. As in our pilot study, we found substantial links between scores on the four goal scales and performance ratings within each classroom, with shrunken $R^2$ values within each class ranging from .09 to .45, with a median of .25. We next examined relationships in the total sample, pooling across instructors. Because each instructor taught a different course, instructors are nested within courses and classrooms, meaning that instructor effects could also reflect variance because of the course or to class-specific factors.

As in our pilot study, we found significant links between scores on the four goal scales and performance ratings obtained at the end of the semester ($R^2 = .10$). We found similar relationships (once again, $R^2$ was significantly different from zero) between goal scale scores obtained at the beginning of the semester and performance ratings collected at the end of the semester ($R^2 = .08$).

Widely followed conventions for interpreting $R^2$ values (e.g., Cohen, 1988; Murphy & Myors, 1998) suggest that values of .25 or larger represent strong effects and values of .10 or larger represent moderate effects. This implies that there is a moderately strong relationship between the goals measured here and ratings.

**Change in Goal Importance**

There were 125 raters who provided goal importance ratings both at the beginning of the semester and at the end of the semester. On the basis of the results of our pilot study, we expected that goal importance ratings collected at the end of the semester would be better predictors of performance ratings than goal ratings collected at the beginning of the semester. The data confirmed our expectation: Incremental $R^2$ values of .09 and .01 were obtained using goal ratings collected at the end of the semester and at the beginning of the semester, respectively.

We tested the hypothesis that changes in goal importance ratings would be related to raters’ evaluations of ratee performance. For example, a rater who believed that his or her instructor performed poorly over the course of the semester might be expected to report more emphasis on weakness goals and less emphasis on strength goals after observing the instructor’s performance than was reported at the beginning of the semester.

Correlations between the measures of goal importance at the beginning and the end of the semester suggest there is only modest stability in the goals raters pursue over time. Eighteen of 19 measures of goal importance obtained at Time 1 were positively correlated with the measures of the importance of the same goals measured at Time 2 (the average of these correlations was .27, with an interquartile range of .22–.29). Time 1–Time 2 correlations for the strength, weakness, fairness, and motivation scales were .30, .60, .27, and .06, respectively. At Time 2, raters were slightly less likely to emphasize motivation goals ($d = -.09$) and somewhat more likely to emphasize weakness, strength, and fairness goals ($d = .09, .08$, and .37, respectively) than at Time 1.

We used hierarchical regression to test the hypothesis that changes in the rated importance of goals from the beginning to the end of the semester would predict end-of-semester teacher ratings. In particular, we first entered instructor effects and measures of goals obtained at the beginning of the semester, and then examined the change in $R^2$ obtained when we added ratings of goal importance obtained at the end of the semester. Increases in $R^2$ for goals at Time 2 indicate that ratings of goal importance obtained at Time 2 contain information not included in Time 1 ratings that is relevant for understanding performance ratings; these $R^2$ increases are usually interpreted as an indication of the relationship between changes in goal importance over time and performance ratings.

As expected, ratings of goal importance collected at the end of the semester accounted for meaningful variance in composite performance ratings. When Time 2 scores on each of the four goal importance scales were added to an equation that already included instructor effects and Time 1 scores on these scales, the change in $R^2$ was .07, which is significant at $p < .05$.

To provide a more detailed test of the hypothesis that changes in goal importance were related to performance ratings, we conducted separate hierarchical regression analyses for each of the 19

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Correlations Among Goal Importance Scales and Performance Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>$M$</td>
</tr>
<tr>
<td>Time 1</td>
<td></td>
</tr>
<tr>
<td>1. Weakness</td>
<td>3.79</td>
</tr>
<tr>
<td>2. Strength</td>
<td>3.90</td>
</tr>
<tr>
<td>3. Fairness</td>
<td>3.88</td>
</tr>
<tr>
<td>4. Motivation</td>
<td>2.93</td>
</tr>
<tr>
<td>Time 2</td>
<td></td>
</tr>
<tr>
<td>5. Weakness</td>
<td>3.84</td>
</tr>
<tr>
<td>6. Strength</td>
<td>3.95</td>
</tr>
<tr>
<td>7. Fairness</td>
<td>4.09</td>
</tr>
<tr>
<td>8. Motivation</td>
<td>2.84</td>
</tr>
<tr>
<td>Performance</td>
<td>4.15</td>
</tr>
</tbody>
</table>

* $p < .05$. 

RESEARCH REPORTS
goal items, each time examining the proportion of variance accounted for by that goal at Time 1 and examining the change in $R^2$ when ratings of goal importance at Time 2 were added to the equation. The results of this analysis are presented in Table 1.

The results presented in Table 1 suggest that changes in goal importance can be useful predictors of performance ratings. For 10 of the 19 goals listed in Table 1, goal importance measured at Time 2 accounted for variance in performance ratings not accounted for by ratings of goal importance at Time 1. Changes in $R^2$ of .10 or larger (indicating moderately large effects) were found for five goals (three weakness and two strength goals).

Discussion

Ratings of goal importance obtained at the beginning of the semester, before students have had a substantial opportunity to observe the instructor’s performance, predicted ratings of instructor performance collected at the end of the semester. Moderately large multiple correlations between goals and performance ratings were found within each class (i.e., in analyses of sets of raters who had all observed essentially the same sample of performance) and were also found in the pooled analysis, in which differences in the mean ratings given to instructors were controlled for prior to estimating the predictive value of goal ratings. These results support our hypothesis that raters who pursue different goals when completing performance ratings do tend to give different ratings, even when they have observed the same performance. Our results are consistent with those reported in our pilot study (which included nearly 900 subjects), but the design of our main study allows us to make stronger inferences about the possible role of rater goals.

The design of our study helps to at least partially rule out a number of plausible alternative explanations for the goal-rating link. In particular, given the lag between collecting goal information and obtaining performance evaluations, reverse causation is an unlikely explanation for our results (i.e., end-of-semester evaluations collected in Study 2 cannot have a direct causal influence on goal ratings collected at the beginning of the semester). The lag between goal measures and ratings also reduces the likelihood that responses to the goal questionnaire prime or affect performance evaluations. There is a gap of several months between the two questionnaires, and during that time, raters have ample opportunities to observe the ratee’s performance. Although goal statements might influence the rater’s perceptions of subsequent performance, it is unlikely that goal ratings collected at the beginning of the semester will have a strong direct effect on end-of-semester performance ratings. Finally, many participants have little or no information at the beginning of the semester about the instructor’s performance, making it unlikely that the performance level of instructors has a substantial influence on goal ratings (however, as we note below, some participants might have at least indirect information about the instructors prior to taking the class). Correlational research, even with longitudinal designs, rarely provides conclusive evidence for causation, but the findings of this study are certainly consistent with a model that assumes that rater goals have an impact on the evaluations they provide.

In traditional supervisory performance appraisals, rating goals are likely to be strongly influenced by raters’ perceptions of the costs and benefits associated with giving high versus low ratings or ratings that discriminate among ratees versus ratings that lump them all together. In our study, ratings were anonymous, and raters had few external incentives to distort their ratings, but we nevertheless found meaningful variation in the goals pursued by raters and consistent links between these goals and performance ratings. Murphy and Cleveland (1995) suggested that external incentives are indeed highly important in influencing rater goals, but they also suggested that the emphasis raters place on specific goals reflects stable individual differences in goal orientations (Dweck & Leggett, 1988, suggested similar individual differences in learning goal orientations). That is, some raters might consistently emphasize strength goals, whereas others consistently emphasize weakness goals, and so on. A small sample of cases from a related study (Murphy, Cleveland, Kinney, Newman, Sin, & Skattebo, 2003) provided evidence suggesting that the ratings of goal importance analyzed here might reflect broad individual differences in the orientations of raters. We were able to identify 17 participants in a related study who completed this same goal importance questionnaire in two different courses, in different semesters, with different instructors. We correlated ratings of the 19 goals across courses and found a median correlation of $r = .38$, with an interquartile range of .18 to .62. It is impossible to make strong claims on the basis of a sample this small, and in any case, our study does not allow us to identify what individual difference variables might contribute to the stability of goal ratings. However, our data are consistent with the hypothesis that rater goals might in part reflect stable orientations on the part of the rater and that raters might consistently pursue a variety of rating goals even in situations where there are weak external incentives to give lenient or distorted ratings or to shade their ratings in any particular way.

Our data also suggest that changes in ratings of goal importance might reflect, at least in part, the rater’s evaluation of the ratee’s performance. For example, goal importance ratings obtained at the end of the semester account for variance in performance ratings not accounted for by goal importance ratings obtained at the beginning of the semester. Specifically, raters who evaluated their instructor more favorably were also more likely to rate the goal of conveying information about strengths more important at the end of the semester than at the beginning of the semester (the correlation between changes in strength ratings and performance ratings was $r = .21$, $p < .05$; we did not find significant correlations between performance ratings and changes in ratings for the other three classes of goals). One interpretation of this finding is that raters adapt their rating strategies as they learn more about the actual performance of the ratee. That is, raters’ whose initial orientation is to place relatively little emphasis on ratee strengths may, when asked to evaluate a truly good performer, come to place more emphasis on conveying information about strengths of that particular ratee. In this study, all instructors were in fact good performers (i.e., their average teacher ratings were all above 4.0 on a 5-point scale), and it makes sense that raters will more strongly emphasize strength goals after observing a semester of good per-

---

1 Teacher ratings are not published at the university in which this research was conducted, and our conversations with undergraduates about their knowledge of instructors suggests that many students come in to the start of a course in the departments studied with few concrete expectations about the instructor’s performance.
Interrater Disagreement as a Function of Rater Goals

Cleveland and Murphy (1992) suggested that an analysis of rater goals could prove useful for understanding so-called “rater errors,” particularly leniency. Analyses based on goal-oriented models might also prove useful for another assessing of psychometric shortcoming in ratings that has received considerable attention in recent years, specifically, low interrater agreement.

In research on performance appraisal, interrater agreement appears to be the exception rather than the norm (Murphy, Cleveland, & Mohler, 2001). Interrater correlations, which reflect rank-order agreement among raters, are rarely much greater than .50 (Viswesvaran, Ones, & Schmidt, 1996), and when performance evaluations are obtained from raters at different organizational levels (e.g., supervisors vs. peers), correlations are even lower (Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988).

Disagreements between raters who evaluate the same set of raters are usually interpreted as evidence of substantial measurement error in performance ratings (Viswesvaran et al., 1996). That is, interrater correlations are sometimes interpreted as reliability coefficients on the basis of the assumption that rank-order agreements between raters reflect the effect of true scores and disagreements about rank-ordering reflect random measurement error. Viswesvaran et al.’s review suggests that the variance accounted for by true scores (approximately 52% of the variance in observed scores) is not much greater than the variance accounted for by measurement errors (according to the classic model of reliability, the remaining 48% of the variance in ratings is accounted for by measurement error). Murphy and DeShon (2000) suggested that interrater correlations are not reliability coefficients, and that the traditional interpretation of why raters agree or disagree in their evaluations (i.e., true scores vs. error) is inconsistent with a substantial body of research on performance evaluation. Rather than dismissing rater disagreements as measurement error, it might be useful to examine in detail why raters who observe the same performance sometimes give similar ratings and sometimes give quite dissimilar ratings.

Our study examines the links between rater goals and performance ratings assigned to instructors in end-of-term teacher ratings. One facet of this particular rating context is that all raters in each class had multiple opportunities to view the same sample of performance, and all raters in each class evaluated the same ratee. The design of these studies, in which there are many raters but only one ratee (i.e., many students rating one instructor) make it impossible to calculate interrater correlations of the sort presented in Viswesvaran et al. (1996), but it is ideal for assessing one of the central assumptions of models that treat interrater disagreements as measurement error.

Interrater disagreements that reflect measurement error are by definition independent of other third variables (this independence assumption is what justifies the correction for attenuation, Lord & Novick, 1968). If interrater disagreements represent rater error, variability in ratings given by judges who have observed the same sample of performance should not, on the whole, be systematically related to rater characteristics such as goals. On the other hand, if the goal-oriented model proposed by Cleveland and Murphy (1992) is correct, information about the goals pursued by different raters who are evaluating the same sample of job performance should account for a substantial portion of variance in their ratings, which means that this variance cannot be treated as random measurement error. Our data support the hypothesis that disagreements among raters who have observed the same sample of performance reflect systematic differences in the raters’ orientation to performance appraisal (e.g., goals); these data do not support models that dismiss interrater disagreements as measurement error.

Limitations

Legitimate questions can be raised about both the internal and the external validity of our study. First, although our design does help to eliminate many alternative explanations for the goal-rating link, it is not possible to obtain the sort of control that might be attained in a true experiment. For example, it is feasible that students know something about the instructor prior to the start of class and that they have at least an impression of the instructor’s skill and teaching style (either on the basis of their previous experience or what they have heard from other students) even before they have observed the instructor’s performance in the class. It is therefore impossible to completely eliminate the chance that students’ prior impressions of the instructor’s effectiveness influence their goal ratings, leading to spurious goal-rating links. It is also possible that some students choose particular classes or instructors at least partly on the basis of their reputations as teachers, and this could influence both the rating goals that are emphasized (e.g., students who seek out instructors who have a reputation for good teaching, lenient grading, etc., might report different rating goals from those who have no choice in instructor or who do not know anything about the instructors). Finally, we have interpreted increments in the variance accounted for by Time 2 ratings not accounted for by Time 1 ratings in terms of the effects of changes in goal importance, but we do not have any direct measures of the intervening processes (e.g., why goals might change) and no way of controlling for plausible alternative explanations for these results (e.g., Time 2 goal ratings might show higher correlations with ratings because of priming or common method effects, as in the pilot study). Although our results are consistent with the hypothesis that changes in goal importance are related to the performance of ratees, these results must be interpreted with even more caution than our own findings, that goals measured at the beginning of the semester predict performance ratings at the end of the semester.

This study relied on student samples and examined anonymous ratings that have no real consequences for the rater and uncertain consequences for the ratee. This is in contrast to more traditional performance appraisals, in which there may be very real consequences for both the rater and the ratee and in which raters must often communicate their evaluations directly with rates (Murphy & Cleveland, 1995). Because of the differences between teacher ratings and supervisory performance appraisals, we must be cautious when drawing conclusions about relationships between goals and ratings on the basis of this study. Organizations are increasingly moving toward 360° rating systems, in which anonymous ratings from many subordinates, peers, and so forth, are pooled and communicated to ratees (Bracken, Timmreck, & Church, 2000),
and it may be that our study has more ecological validity for drawing conclusions about anonymous rating systems than about ones in which ratings have a direct impact on both the rater and the ratee. However, even in these contexts, there are good reasons for caution in drawing inferences from our results. In particular, we relied on Kerst’s (1993) study of managers in developing our rating goal questionnaire. It is possible that as we move away from the context of traditional supervisory ratings, the types of goals that are most important to raters change.

As we noted earlier, one important difference between the context in which ratings were obtained in this study and more typical performance appraisals is that in our study, raters had little prior knowledge of and often no expectation of future interactions with the ratee (i.e., their course instructor). In typical performance appraisals, raters develop clear expectations about ratee performance, which in turn are likely to influence both rating goals and future performance ratings. Over time, it is reasonable to expect that raters will adjust their rating goals in response to the success or failure of their previous performance appraisals (e.g., a rater who hopes to improve performance by giving high ratings might stop inflating ratings if no changes in performance are observed). Therefore, unlike the present study, the process of forming and applying rater goals in performance appraisal is likely to be a dynamic one. Our data do suggest that changes in goals are related to performance ratings, but field studies of the development and modification of rating goals would provide a clearer sense of how rating goals and their influence on ratings change over time.

Finally, our studies do not provide any information about the origin or development of rater goals or about the variables that influence the goals different raters pursue. Murphy and Cleveland (1995) offered a number of hypotheses about the origin and development of rater goals, but these have not been empirically tested to date. Our study does not answer the question of why raters (especially those with few external incentives to distort their ratings) might pursue different rating goals, but it does suggest that the goals raters pursue are systematically related to the ratings they give.

References


Received May 9, 2002
Revision received April 29, 2003
Accepted May 8, 2003