Conscientiousness and Agreeableness as Predictors of Rating Leniency

H. John Bernardin and Donna K. Cooke  
Florida Atlantic University

Peter Villanueva  
Appalachian State University

The authors studied the relationship between the Big Five personality factors and rating elevation among 111 students making peer evaluations. It was hypothesized that Conscientiousness (C) scores would be negatively correlated with rating level and that Agreeableness (A) scores would be positively correlated with rating level. We further predicted that individuals who were low in C and high in A would produce the most elevated ratings. As predicted, A scores were positively related to rating level (r = .33, p < .01) and C scores were negatively correlated with rating level (r = -.37, p < .01). Using the strong hypothesis test (P. Bobko, 1986), the authors found that ratings by low C and high A individuals were more elevated than all other groups of students combined (p < .01).

Rating elevation is apparent when ratings are largely restricted to the positive end of the continuum on a rating scale. Barrett (1966) considered the problem endemic to "virtually every rating program. When a program is initiated, more than half the people are given ratings above average and the proportion of high-rated people grows until only the obvious misfits fail to make the top grades" (p. 23). Several negative consequences may befall operational-performance management systems characterized by elevated ratings. Elevated ratings reduce the size of the available funds used to recognize and reward performance and therefore may give rise to employee dissatisfaction with both performance management and reward systems. Elevated ratings also make it difficult to substantiate termination decisions, not only to the courts if decisions are contested, but also to remaining employees who may then question the fairness and meaningfulness of performance ratings. Elevated ratings may also present a Pollyannan view of organizational effectiveness and result in a complacent organizational assessment.

Bretz, Milovich, and Read (1992) reported that 77% of sampled companies believe that lenient appraisals jeopardized the validity of their appraisal systems. Earlier, Bernardin and Villanova (1986) found that a majority of both supervisors and administrators of appraisal systems indicated that the desire to avoid confrontations with subordinates was a significant source of appraisal inaccuracy in their organization. In the same study, a majority of subordinates reported that raters rate higher than deserved to please certain employees. Murphy and Cleveland (1991) wrote, "it is well known that ratings are skewed in a positive direction and that many rates receive higher ratings than they deserve" (p. 289).

Despite the pervasiveness and acknowledged seriousness of the problem, little is known about the individual versus situational determinants of rating elevation and the boundary conditions that influence such determinants. The limited research on the subject is beset by a host of methodological problems, such as reliance on a single rating observation (Murphy & Cleveland, 1991) and research designs that ignore potential boundary variables that may be necessary to elicit detectable levels of elevation (e.g., rater accountability, personal relationships between raters and ratees, purposes for data collection) that operate in actual appraisal situations. Also, too few studies examining the prediction of elevated ratings chose individual difference measures on the basis of something other than convenience (Villanova, Bernardin, Dahms, & Sins, 1993).

Like leniency, rating elevation seems almost immune to any means of appropriate and acceptable control. For example, complex scoring systems such as binary scoring (Bass, 1956), mean transposition (Bernardin & Beatty, 1984), or statistical control (e.g., Stricker, Rock, Burton, Muraki, & Jirele, 1994) usually result in user resistance to the scoring systems' application as well as confusion as to what meaning is to be attributed to the transformed scores.

Recent research has shown that rating elevation is a relatively stable characteristic on the part of the rater (Borman & Hallam, 1991; Kane, Bernardin, Villanova, & Peyrefitte, 1995; Villanova et al., 1993). For example, Kane et al., (1995) found a mean stability coefficient of .48 across three studies. Villanova et al. (1993) reported a mean stability coefficient of .63 for students evaluating their peers on group projects. These findings suggest that rating elevation might be predicted using measures of individual differences.

The purpose of this research was to attempt to predict rater elevation using individual-difference variables. Specifically, we used the five-factor model (FFM) of personality to predict mean ratings by raters (Costa & McCrae, 1992; Digman, 1990). Although controversy remains (e.g., Hough & Schneider, 1996), the FFM structure has been gaining widespread acceptance by personality researchers and has greatly influenced the research on individual differences (Barrick & Mount, 1991; Goldberg, 1993; Sal-
ELEVATED RATINGS AND THE BIG FIVE

There appears to be a growing accord about a general framework for the taxonomy of personality traits. Most researchers have found five replicable factors (e.g., Borgatta, 1964; Digman & Inouye, 1986; Digman & Takeken-Chock, 1981; Fiske, 1949; Goldberg, 1990, 1992; McCrae & Costa, 1985, 1987; Norman, 1963). The reduction to five factors is not intended to reduce the measurement of personality to five factors, but rather to provide a scientific framework to organize the many individual-difference measures (Goldberg, 1993). Extensive reviews of these studies have also been conducted (e.g., Digman, 1990; McCrae & John, 1992; Wiggins & Pincus, 1992) providing further confirmation of the existence of the five factors.

Among these Big Five factors, Conscientiousness (C) and Agreeableness (A) seem particularly relevant for explaining rating elevation. Individuals scoring high on C strive for excellence, have high performance standards, and set hard-to-accomplish goals (Costa & McCrae, 1992). Thus, the ratings provided by high-C individuals, as careful and thorough people, would be less prone to elevation. More agreeable individuals are more trusting, sympathetic, cooperative, and polite. However, high scores on A may also be dysfunctional. High-A individuals may be more dependent and self-effacing, to a fault. A strong orientation toward agreement and desire for social approval in situations that require assertiveness and independence for successful resolution would tend to increase the desire to avoid social conflict and potentially contribute to rating elevation. Thus, people high in A are likely to produce more elevated ratings. We further predict that scores on A interact with scores on C to augment elevation. Specifically, raters who are low in C and high in A generate higher rating levels than all other raters.

A previous attempt to predict rating elevation using constructs similar to the FFM did not support the predictive value of these constructs. Villanovana, Bernardin, and Ross (1997) found that scores for the 16 personality factor second-order factor used to approximate A (dependence) correlated positively with peer rating level (r = .33, p < .01), but that scores for the second-order C factor were not related to peer rating level (r = .04, ns). Moreover, whereas the relationship between A scores and rating level increased as a function of criterion aggregation (r = .16, for two ratings; r = .23, for three ratings; and r = .33, for four ratings), the pattern for C scores was reversed such that increased numbers of ratings resulted in consistently smaller relationships between the C scores and rating level aggregation (r = -.16, for two ratings; r = -.13, for three ratings; and r = -.04, for four ratings). In this study, we sought to conduct a better test of the hypothesized relationships between personality and rating level by using a higher fidelity measure of the Big Five without resorting to second-order factor analysis of first-order factors as was used by Villanovana et al. (1997).

Method

Overview

One hundred eleven students completed the NEO Five Factor Inventory (NEO-FFI) (Costa & McCrae, 1992) and participated in group exercises involving a variety of human resource management problems. The students each evaluated each other in terms of how they performed in the group exercises. The primary predictors of interest in this study were the A and C scores as estimated from the NEO-FFI. The primary dependent variable was the average rating each student provided to the other students in the exercises. In addition, averaged professor ratings for student performance were available as a true score estimate.

Participants

Students (N = 111) participated in the study as a requirement for an undergraduate course in human resource management. All students took one of five sections of the same undergraduate class, all of which were taught by the same instructor across a 2-year period. There were no significant differences among the characteristics of the five classes on any critical variables under study, including scores on any of the FFM factors, the mean peer rating, student demographics, or grade point average. The range in class size was 20 to 32 students.

Procedure

Students participated in a peer review process that was one of the course requirements. Each student participated in from four to six different group exercises in which group membership ranged from 4 to 8 students and group membership for a particular exercise was randomly assigned (Bernardin & Russell, 1993, Appendix B). The exercises involved an individual and a group component in which each student had to prepare a written response prior to class and take a definitive position on an issue related to human resource management. In designated groups, all members had to review the written responses of the other members and then attempt to reach consensus on the issue. All exercises came from a textbook on human resource management (Bernardin & Russell, 1993). The exercises included a case on a possible age discrimination suit, a staffing exercise, a recruitment and employee turnover problem, consideration of a pay equity policy for state employees, a union organizing case, and a performance appraisal exercise.

Variables

NEO-FFI. Students completed the NEO-FFI (Costa & McCrae, 1992) as part of the course requirement. Students were requested but not required to provide their names (data used here are from students who volunteered their names). A total of 16 students did not volunteer their names for the study. The NEO-FFI is a 60-item version of the NEO Personality Inventory-Revised that provides a "brief, comprehensive measure of five domains of personality" (Costa & McCrae, 1992, p. 11). Costa and McCrae (1992) reported that the coefficient alphas for the NEO-FFI subscales based on a sample of 1,539 adults were .68 for Agreeableness, .77 for Extraversion, .73 for Openness, .81 for Conscientiousness, and .86 for Neuroticism. The instrument was completed in the first 2 weeks of the semester.

Peer ratings. After participating in a group exercise, each student made peer evaluations of group members on eight performance competencies (see Bernardin & Russell, 1993, p. 676). The students were required to submit their peer ratings as part of the course assignment. Students were informed that peer-rating data were to be used in partial determination of the class participation grade for each student. Each rater received a feedback report at the conclusion of each exercise that included the ratings identified by rater. Raters were also asked to make open-ended comments about each peer group member. Students were also told that the manner in which they conducted their peer appraisals would be evaluated in the context of the directions presented in the text (Bernardin & Russell, 1993, Appendix B) and that the instructor would evaluate each student's participation in the peer evaluation process.

The competencies rated were analytic thinking, behavioral flexibility, decision making, leadership, planning and organizing, personal impact, oral communication, and written communication. Peer ratings were made on eight dimensions using a 7-point scale ranging from 1 (poor) to 7
Table 1

<table>
<thead>
<tr>
<th>Personality trait</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Professor’s rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Neuroticism</td>
<td></td>
<td></td>
<td>.36</td>
<td>.02</td>
<td>.01</td>
<td>.11</td>
<td>.12 .03</td>
</tr>
<tr>
<td>2. Extraversion</td>
<td></td>
<td></td>
<td></td>
<td>.19</td>
<td>.08</td>
<td>.40</td>
<td>.13 .14</td>
</tr>
<tr>
<td>3. Openness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.07</td>
<td>.07</td>
<td>.13 .08</td>
</tr>
<tr>
<td>4. Agreeableness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.06</td>
<td>.33 .01</td>
</tr>
<tr>
<td>5. Conscientiousness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.37 .00</td>
</tr>
<tr>
<td>6. Average rating</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19.11 31.83 28.12 33.55 37.68 5.51 4.42</td>
</tr>
<tr>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.94 6.23 6.73 6.23 6.93 0.66 0.31</td>
</tr>
</tbody>
</table>

Note. N = 111. Coefficients greater than .19 are statistically significant at the .05 level, two-tailed.

(outstanding): The mean peer rating across all exercises and dimensions was taken as the measure of rating level for each rater (see Table 1, average rating). The mean number of exercises in which a student participated and rated peers was 4.6 per rater, and the mean number of total rates per rater was 26.2 (SD = 8.1).

Professor ratings. The class professor also provided ratings of student exercise performance. The professor reviewed students’ written products and provided observation and supervision of group performance as each group completed the in-class portion of the assignments. The professor rating used here is the average rating by the professor of each ratee common to each rater. The professor ratings perhaps better reflect the analytical thinking and written-communication dimensions of student exercise performance. Although not as comprehensive as the peer ratings, the professor ratings serve as a criterion to be compared to the peer ratings.

Results

Table 1 reports descriptive data, including correlations between the Big Five factors and the average peer rating for the 111 participants who provided complete study data. As shown in Table 1, the A and C scores of the Big Five were independent, whereas Extraversion scores were negatively correlated with Neuroticism scores (r = -.36) and positively correlated with both Openness (r = .19) and Conscientiousness scores (r = .40). One of the means were significantly different from the means reported for college students in the manual for the NEO-FFI. Also, there were no significant differences (p < .05) in FFM scores between respondents who volunteered their names and those who did not.

As we hypothesized, the relationship between A and rating level was positive and statistically significant (r = .33, p < .01). In addition, the statistically significant correlation between C and ratings was consistent with the predicted negative direction (r = -.37, p < .01). None of the other Big Five factors was statistically related to peer ratings. The raters’ FFM scale scores did not correlate significantly with professor ratings. Average peer ratings by raters were largely unrelated to professor ratings of the same students (r = .09, ns).

Analysis of the hypothesized A × C interaction proceeded in two stages. First, we analyzed the average peer-ratings data using a hierarchical moderated regression analysis and then followed up using an alternative method of analyzing a priori ordinal interactions proposed by Bobko (1986). The moderated regression analysis revealed that the main effects for A, β = .33, t(107) = 3.74, p < .01, and C, β = -.36, t(107) = -4.26, p < .01, were both statistically significant. The A × C interaction was not statistically significant, t = 0.04, ns. We also conducted the strong hypothesis test for an ordinal interaction described by Bobko (1986). We posited that the outlying group mean that would be responsible for the predicted interaction would come from the high-A–low-C cell. We divided our sample into four groups on the basis of their A and C scores relative to published norms for college students. Because the three group means were found to be largely equivalent, F(2, 88) = 2.2, p > .10, we proceeded to the two-group comparison through a t test. This t test was statistically significant, t(109) = 2.45, p < .05, and indicated that the high-A–low-C mean performance rating (M = 5.83) was higher than the mean from the pooled cells (M = 5.43) to which it was compared and higher than all remaining cells as well (mean low A and low C = 5.48; mean low A and high C = 5.25; mean high A and high C = 5.57).

We then repeated these analyses with the difference score computed by subtracting the average professor’s rating from the average peer rating for each rater. The professor ratings were lower than those provided by the student raters of their peers (mean professor ratings = 4.42 vs. mean student rating = 5.51), and this difference was statistically significant according to a paired samples t test, t(110) = 16.2, p < .01. The hierarchical moderated regression conducted on the difference scores computed between average peer ratings and professor ratings largely replicated the analysis reported above. That is, like the peer-ratings data, both the A and C main effects were statistically significant, β = .31, t(107) = 3.67, p < .01, and β = -.34, t(107) = -3.94, p < .01, respectively, and the interaction of A × C was again not statistically significant, t(109) = 0.61, ns. However, for these data, the strong hypothesis test for ordinal interactions did not replicate the peer-ratings results. The mean of the high-A–low-C cell did not differ significantly, t(109) = 0.87, ns, from the mean calculated from the pooled difference scores across the remaining three groups.

Discussion

The results of this study suggest that rating elevation may be predictable from scores on two factors of the FFM. Results of the hierarchical moderated regression analysis indicated that, as expected, individuals scoring higher on the A factor tended to provide more elevated ratings of peer performance on group exercises, whereas individuals scoring higher on the C factor rated peers lower. It appears that in the context of peer ratings, scores from the FFM might be useful for identifying potential elevation in performance ratings. Also, the results of the strong hypothesis test (Bobko, 1986) suggested that individuals with high A and low C scores produced the most elevated ratings observed in this study.
Boundary Conditions

This study used undergraduate students as raters and ratees. The extent to which our results generalize to other appraisal situations is unknown. However, this rating situation controlled important boundary variables characteristic of a modal criterion model (e.g., Bernardin & Villanova, 1986). For example, raters were identified in the rating process, and the ratings had administrative significance.

Applied Implications

The potential applied implications of these findings include the use of personality inventories or assessment for supervisor selection in those instances where supervisory behavior in the performance appraisal process may be considered an important issue. Alternatively, rater scores on the FFM might also be used in a diagnostic manner to identify raters who may need self-efficacy training (Bandura, 1986) to reduce their tendency to rate high. Previous research has shown that attitudes toward appraisal processes are predictive of elevation (Villanova et al., 1993), and diagnosis of rating elevation may be enhanced using measures of both dispositional constructs, such as the FFM factors, in combination with measures of attitudinal constructs such as the Performance Appraisal Discomfort Scale (Villanova et al., 1993).

Future Research

Kane et al. (1995) advanced a conditional view of dispositional constructs to explain elevation. This view maintains that some traits are more or less predictive of behavior under specific conditions. For example, perhaps we found A and C to be predictive of elevation in this study because the context possessed specific features that were relevant to these traits, such as accountability, the potential for future interaction, and the administrative significance of the ratings. It may be that if any one of these features were absent, the predictability of rating level from these trait scores might be attenuated. Future research should investigate the boundary conditions that influence trait—behavior relationships in the appraisal context, as well whether the predictability of rating elevation may be enhanced by measuring more specific facets of the Big Five factors.

It may also be that personality may interact to influence ratings when combined with different treatments. For example, one reviewer raised the issue of how A and C might interact with rater training. It could be that individuals higher in A and C respond more favorably to rater training, such as frame of reference training (Sulsky & Day, 1992). If such is the case, then the effect for C would likely be magnified and the effect for A muted or even reversed if more agreeable individuals seek approval primarily from those authorized to provide training.

Finally, this study involved an issue common to criterion research—the appropriate labeling of constructs (cf. Austin, Villanova, Kane, & Bernardin, 1991). Because we could not strongly infer that these ratings reflected leniency per se (defined here as some positive deviation from a true score), we chose to avoid this term. Rather, we do believe the current data are demonstrative of high rating levels as indicated by the mean peer ratings relative to the absolute scale values and also relative to the fallible true score estimate represented in the professor ratings. Because the peer ratings deviate positively from the professor ratings, we believe these ratings share some common conceptual basis with leniency.

Just the same, rating level and leniency could conceptually be independent constructs; they need not share the redundancy that seems intuitive and plausible. Figure 1 illustrates instances where leniency and rating elevation may be common and when they may diverge. A conventional example of when leniency and rating elevation co-occur is easily understood: A mean rating of 4.5 on a 5-point scale may be elevated relative to the scale midpoint (3) and also show leniency relative to a true score estimate of 4.0 (as described in Cell D). Less easily imagined scenarios are those where leniency and rating elevation do not co-occur, as shown in Cells B and C. The relative base rate of such occurrences as those represented in Cells B and C is unknown but very likely rare, relative to the occurrences illustrated in Cells A and D. Although it appears that leniency and rating elevation are very probably co-occurring most of the time, there may be rare instances where they diverge. Precise labeling of rating phenomena such as leniency may require more effort than heretofore expected.

References


Received April 21, 1998
Revision received May 3, 1999
Accepted May 7, 1999