Effects of Ratee Task Performance and Interpersonal Factors on Supervisor and Peer Performance Ratings

Walter C. Borman
University of South Florida and PDRI, Inc.

Leonard A. White
U.S. Army Research Institute

David W. Dorsey
University of South Florida

The study examines the effects of a wide array of rater–ratee relationship and ratee-characteristic variables on supervisor and peer job-performance ratings. Interpersonal ratings, job performance ratings, and ratee scores on ability, job knowledge, and technical proficiency were available for 493–631 first-tour U.S. Army soldiers. Results of supervisor and peer ratings-path models showed ratee ability, knowledge, and proficiency accounted for 13% of the variance in supervisor performance ratings and 7% for the peer ratings. Among the interpersonal variables, ratee dependability had the strongest effect for both models. Ratee friendliness and likability had little effect on the performance ratings. Inclusion of the interpersonal factors increased the variance accounted for in the ratings to 28% and 19%, respectively. Discussion focuses on the relative contribution of ratee technical and contextual performance to raters' judgments.

Performance ratings continue to be the most often used criterion measure for personnel research applications (e.g., Murphy & Cleveland, 1991). Research on ratings also remains popular (e.g., Borman, 1991), with the motivation great to reduce rating errors and increase the accuracy of performance evaluations. Streams of research have included rating format studies (e.g., Bernardin, 1977), rater training initiatives (e.g., Pulakos, 1984; D. E. Smith, 1986), and cognitively oriented studies of the performance rating process (e.g., Cardy & Dobbins, 1994; Murphy & Cleveland, 1991; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982).

Walter C. Borman, Department of Psychology, University of South Florida, and PDRI, Inc.; Leonard A. White, U.S. Army Research Institute, Alexandria, Virginia; David W. Dorsey, Department of Psychology, University of South Florida.

This research was performed under Project A, Contract MDA903-82-C-0531 and the Career Forces Contract MDA903-89-C-0202 for the U.S. Army Research Institute (ARI), John P. Campbell, Principal Investigator. However, opinions and conclusions presented here do not necessarily reflect those of ARI, the U.S. Army, or any other agency of the U.S. government.

We thank the following persons for their help at one stage or another of the research: Elaine Pulakos, Scott Oppler, John Campbell, Mike Rumsey, Mary Ann Hanson, Mike Coover, Jim Harris, and Joel Leikowitz.

Correspondence concerning this article should be addressed to Walter C. Borman, Department of Psychology, BEH 339, University of South Florida, Tampa, Florida 33620-8200.

One direction for rating-process research has involved field investigations of the factors and cues raters use in making evaluations of work performance. Policy capturing (e.g., Hobson, Mendel, & Gibson, 1981; Zedeck & Kafry, 1977), a lens-model conceptualization (Schmitt, Noe, & Gottchalk, 1986), and performance-rating models using path analysis or structural models have been used to assess the influence of various factors on, especially, overall supervisory performance ratings.

As an example of this latter approach, Hunter (1983) conducted a path analysis with data from 4 civilian and 10 military studies to evaluate causal relationships between cognitive ability, job knowledge (paper-and-pencil test scores), task proficiency (work-sample scores), and overall supervisory performance ratings. Results showed that ratee job knowledge had the largest direct path to the ratings, approximately three times as large as the ratee task proficiency → ratings path. Cognitive ability had an indirect influence on the ratings, primarily through its effect on knowledge acquisition. Schmidt, Hunter, and Outerbridge (1986) extended the Hunter model to include job experience as an additional variable. Results for the ability, job knowledge, and task proficiency effects on ratings were similar to Hunter's findings. Experience had a moderate effect on ratings, primarily through its influence on job knowledge.

More recently, Borman, White, Pulakos, and Oppler (1991) tested an expanded version of the Hunter (1983)
and Schmidt et al. (1986) models. We used LISREL VII on Army Project A (Campbell, 1985; Eaton, Goer, Harris, & Zook, 1984) data, which included the Hunter variables plus two ratee temperament constructs (Achievement and Dependability), number of awards, and number of disciplinary infractions. Results showed that technical proficiency and ratee disciplinary problem behavior had the strongest direct effects on the supervisory ratings. This expanded model fit the data very well and accounted for approximately twice the variance in ratings in this study than did Hunter's variables alone.

By including ratee temperament variables in the model, Borman et al. (1991) began to respond to Guion's (1983) challenge to include more ratee personal characteristics and rater–ratee relationship factors in these models. Guion's point was that these kinds of interpersonal variables are likely to influence ratings over and above the ratees' ability and competence on the job (e.g., job knowledge and technical proficiency). The research by Borman et al. confirmed this hypothesis but was limited in that only two such variables (plus awards and disciplinary cases) were included. The present study represents an attempt to respond further to Guion's (1983) suggestion to study the effects on ratings of additional rater–ratee relationships and ratee personal-characteristic variables. Thus, the objective of this research is to investigate the effects of several characteristics of ratees on rater judgments about overall ratee performance, with an emphasis on the interpersonal domain.

This research can also be framed in the "cold" versus "hot" processing literature. The cognitive, information-processing, theoretical approach to performance appraisal has emphasized rational models of the appraisal process (e.g., De Nisi, Cafferty, & Meglino, 1984; Landy & Farr, 1980). Ratee personal characteristics and rater biases are considered in these models, but the central focus is on the observation-encoding-storage-retrieval-judgment steps of the rating process.

Somewhat in contrast, Zajonc (1980) argued that "hot cognitions" can play a very important role in person perceptions. By "hot cognitions," Zajonc meant affective reactions that may occur independent of the cognitive operations usually thought to precede in time affect toward these stimuli. He reviewed research suggesting that affective reactions can occur without cognitive processing steps being evident and that judgments arising out of these reactions can be made with greater confidence than cognitive-driven judgments. In relation to judgments about others' job performance, this research suggests that affect (e.g., feelings of liking or disliking) may be influential beyond the cognitive steps of observation, encoding, and so forth in the performance rating process.

Some performance appraisal literature exists that demonstrates the potential role of affect and liking in performance judgments. Williams (1986) found that negative affect had a significant negative effect on performance evaluations of hypothetical employees. Tsui and Barry (1986) demonstrated a correlation between liking and a leniency effect, with better-liked ratees rated as more effective performers. However, the direction of causation in this study was difficult to determine. High performance could have positively influenced liking, and liking could have led to higher performance ratings. In another study with 24 undergraduate raters as participants, Harris and Sackett (1988) used a policy-capturing analysis to show that the liking of major league baseball players contributed independent variance to performance ratings of the players for about one-half of the student participants. Finally, Dobbins and Russell (1986), in both a lab and a field study, found raters reporting a higher likelihood of punishing subordinates for poor performance if they were less well liked; Wayne and Ferris (1990) obtained a significant liking → performance rating path from path analyses done on both a student sample and a bank employee sample; and Cardy and Dobbins (1986) found liking of ratees to adversely affect raters' accuracy in evaluating performance. Accordingly, ratee likability was seen as an important factor to consider in the present research.

In a related study with state patrol officers, Gilbert and Whiteside (1988) identified several ratee characteristics as significant correlates of positive performance evaluations. One of these characteristics was ratee dependability. This is interesting because our previous research (Borman et al., 1991), where ratee dependability was measured by a self-report temperament scale, showed it to have both a direct and an indirect effect on supervisory ratings. Also, Barrick, Mount, and Strauss (1993) found that ratee conscientiousness substantially influences performance judgments, both directly and through its effect on goal-setting by ratees. The Barrick et al. measure of conscientiousness contains as elements dependability and achievement-orientation. Accordingly, dependability was a second ratee characteristic variable considered for the present research.

In addition to including ratee likability and dependability in the study as constructs hypothesized to influence performance judgments, our intention was to identify additional rater–ratee-relationship and ratee-characteristic variables by interviewing a sample of raters and asking them to describe the kinds of interpersonal factors that might influence their or others' job performance ratings. The idea was to represent, in a final set, a relatively comprehensive array of interpersonal factors that have potential for influencing performance ratings.

Finally, all of the rating models discussed previously examined factors and cues supervisor raters use in making performance judgments. Thus, for example, peer rat-
ings, used in personnel research applications (e.g., Kane & Lawler, 1978) and to aid in making personnel decisions (e.g., Edwards, Borman, & Sproul, 1985), have not been studied in this context. And yet, there is reason to believe that peers might use different patterns of factors and cues in making judgments of performance. For example, Klkoski and London (1974) argued from their factor analysis results of self, peer, and supervisor ratings that raters from different organizational levels use different factors in making ratings. Borman (1974) found that supervisor and peer raters of clerical workers generated different dimensions in behaviorally anchored rating scale workshops. The evidence is mixed on this topic (cf. Harris & Schaubroeck, 1988; Tsui & Ohlott, 1988), but peers generally see different samples of ratee behavior and have different organizational roles, vis-à-vis their coworkers, in comparison with supervisors. Accordingly, we studied peer-rating models, as well as supervisor-rating models, in the research.

Thus, the current research extends previous work in three important ways. First, the number and type of variables related to the rater–ratee relationship and ratee personal characteristics was greatly expanded. Second, the focus of the study is on both supervisor and peer ratings. Third, we avoided a potential confound in this kind of study. Previous work has sometimes had the same rater making both the ratings on the interpersonal variables and on performance. Halo and differential leniency across raters may operate to inflate correlations between those ratings under those conditions. In this study, for the supervisor-rating model, different supervisors provided ratings on the interpersonal factors and the performance ratings. Similarly, for the peer-rating model, one peer for each ratee made the performance ratings, and the mean of all other peers' ratings on the interpersonal factors was used.

Method

Participants

First-tour U.S. Army soldiers (N = 805) participated in the research. Almost all had spent 36 to 40 months in the Army; 28% were African-American, 3% were Hispanic, and 12% were female. The soldiers served in five different military occupational specialties (or jobs) as follows: 172 infantrymen, 168 armor crewmembers, 148 radio teletype operators, 156 light wheel vehicle mechanics, and 161 medical specialists.

Of the 805 participants, a total of 493 were rated by two or more supervisors, and 631 were rated by at least two peers. Supervisors were typically first-line (e.g., squad leader) and second-line (e.g., platoon sergeant). For both supervisors and peers, raters had worked with their ratee for at least 1 month and reported they were familiar with his or her performance.

Instruments

Available for the ratees were scores on a cognitive ability test, work-sample and job-knowledge tests, rating scales measuring the supervisor-and-peer interpersonal factor, and rating scales measuring job performance.

Cognitive ability test. Before entering the military, all ratees had completed the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB is composed of 10 subtests and is used for selection and classification. The Armed Forces Qualifications Test (AFQT) composite of four ASVAB subtests (Word Knowledge and Paragraph Comprehension, Arithmetic Reasoning, and Mathematics Knowledge) has been reviewed as a good measure of cognitive ability, or g (Murphy, 1984), and was therefore used as the cognitive-ability measure in the present research.

Job knowledge tests. For each of the five jobs, 30 important knowledge areas were identified in job analysis work, and items were written in consultation with subject-matter experts to tap these areas (Campbell, Campbell, Rumsey, & Edwards, 1986). Fifteen of these knowledge areas corresponded to the task-proficiency measures (described later) developed for each job. The knowledge tests were multiple choice, and each test contained 150 to 200 items. For each soldier, the overall job-knowledge test score was the percentage of correct answers.

Hands-on task proficiency measures. For each job, 15 critical tasks representative of the entire task domain were identified as targets for test development work (e.g., for the job of medical specialist, tasks included “assemble a needle and syringe” and “initiate an intravenous infusion”). Task-proficiency tests were prepared for each of the 15 target tasks (Campbell et al., 1986). Each task had several performance steps, each of which was scored as passed or failed. A proportion-passed score was derived for each soldier on each task, and these proportions were averaged across tasks to yield an overall task-proficiency test score.

Interpersonal factors rating scales. The literature review described briefly above suggests that ratee likability and dependability should be considered as interpersonal constructs to include in the research. In addition, we interviewed 25 first-line supervisors in a variety of Army jobs and asked them about the rater–ratee relationship and ratee characteristics factors they, other supervisors, or peers of first-tour soldiers might use in judging the performance and overall value to the unit of these soldiers. Several factors were identified in the interviews, and each interpersonal factor dimension was carefully defined with the help of four to five of these interviewees. Dimension labels and scale definitions appear in Table 1. As can be seen, 11 of the dimensions are in common across the supervisor and peer groups, and 6 additional dimensions were identified for peer raters.

To evaluate the dimensionality of these interpersonal dimensions, the supervisor and peer ratings on the dimensions were intercorrelated, and each correlation matrix (separately) was factored using the principal-axis method with subsequent promax rotation. Results for the supervisor interpersonal ratings matrix suggested three factors, with the following dimensions loading primarily on them: Dependable/Trustworthy—Count on for Backup, Trust and Depend On; Friendly/Easy to Get
### Table 1
**Interpersonal Factor Dimensions and Definitions**

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Sense of Humor</td>
<td>Enjoys and appreciates funny and amusing things.</td>
</tr>
<tr>
<td>2. Irritable</td>
<td>Easily annoyed or angered.</td>
</tr>
<tr>
<td>3. Know-It-All</td>
<td>Acts as if he or she knows everything about every topic.</td>
</tr>
<tr>
<td>4. Friendly</td>
<td>Often seeks others' company; enjoys being with others.</td>
</tr>
<tr>
<td>5. Socially Skilled</td>
<td>Smooth and effective in handling relationships with other people.</td>
</tr>
<tr>
<td>6. Moody</td>
<td>Tends to have large mood changes over time.</td>
</tr>
<tr>
<td>7. Likable</td>
<td>Easily liked and pleasant to be with.</td>
</tr>
<tr>
<td>8. Rude</td>
<td>Impolite and ill-mannered.</td>
</tr>
<tr>
<td>9. Complaining</td>
<td>Gripes a lot and often complains about his or her situation, treatment, etc.</td>
</tr>
<tr>
<td>10. Count on for Backup</td>
<td>I can count on this soldier to back me up if I really need it.</td>
</tr>
<tr>
<td>11. Trust and Depend On</td>
<td>I can trust and depend on this soldier.</td>
</tr>
<tr>
<td>12. Generous</td>
<td>Giving and willing to share with others.</td>
</tr>
<tr>
<td>13. Mean</td>
<td>Is nasty to other people.</td>
</tr>
<tr>
<td>14. Bootlicker</td>
<td>Tries to flatter the supervisor to get ahead.</td>
</tr>
<tr>
<td>15. Boastful</td>
<td>Brags a lot.</td>
</tr>
<tr>
<td>16. Discuss Personal Problems</td>
<td>If I had a personal problem, I would feel free to discuss it with this soldier.</td>
</tr>
<tr>
<td>17. Provide Good Advice</td>
<td>If I had trouble performing a task, I could go to this soldier for good technical advice.</td>
</tr>
</tbody>
</table>

**Note.** Dimensions 10, 11, 16, and 17 had 5-point scales ranging from 1 = *strongly disagree* to 5 = *strongly agree*; all other dimensions had 5-point scales ranging from 1 = *extremely inaccurate* to 5 = *extremely accurate.*

Along With—Sense of Humor, Friendly, Social Skills, Likable; Obnoxious/Complaining—Irritable, Moody, Complaining, Know-It-All, Rude. Analysis results for the interpersonal-ratings matrix for peers showed four factors, with the following dimensions loading on them: Dependable/Trustworthy—Count on for Backup, Trust and Depend On, Discuss Personal Problems, Provide Good Advice; Friendly/Easy to Get Along With—Generous, Sense of Humor, Friendly, Social Skills, Likable; Obnoxious/Nasty—Irritable, Mean, Moody, Complaining, Rude; Show-off—Know-It-All, Boastful, Bootlicker. The Dependability/Trustworthy factors appeared to reflect well the dependable variable emerging from the literature review. Unit weighted composites of the standardized variables were formed for each of these factors. Coefficients alpha reliabilities for the three supervisor composites were .84, .82, and .81. For the four peer composites, these reliabilities were .82, .80, .82 and .75. The corresponding one-rater interrater reliabilities for the supervisor composites were .57, .45, and .29; for the peers, the interrater reliabilities were .33, .38, .50, and .47.

**Performance rating scales.** The performance-ratings scales developed were intended to be appropriate for assessing effectiveness in any first-term Army job and, accordingly, were referred to as Army-wide scales. A variant of the behaviorally anchored rating scale method (BARS; P. C. Smith & Kendall, 1963) was used to develop the scales. Seventy-seven commissioned and noncommissioned officers generated more than 1,100 incidents describing first-term performance; analysis of the content of these incidents and a subsequent retraining step resulted in 10 behaviorally based rating dimensions.

For the research, a unit-weighted composite of those 10 (standardized) ratings was used. We believe this approach is justified, as the dimensions are a result of a thorough critical-incidents job analysis and therefore should reflect all important performance requirements for these jobs. The coefficient alphas for the composites were .65 for the supervisor ratings and .51 for the peer ratings. One-rater interrater reliabilities for the composites were .48 and .35, respectively. More details regarding the development of the rating scales can be found in reports by Borman, Motowidlo, Rose, and Hanser (1987) and Pulakos and Borman (1986).

**Procedures**

Supervisors and peers of the first-tour ratees were trained to use the BARS. With reference to the rater training literature (e.g., Bernardi & Buckley, 1981; Pulakos, 1984), the training can be characterized as a combination of psychometric error and frame-of-reference program. In groups of 3 to 20, supervisors and peers were trained, completed their BARS ratings, and then were administered the interpersonal-factors scales. Each supervisor rated an average of four subordinates (range = 1–8); each peer rated an average of five raters (range = 1–7).

The first-tour soldiers (ratees) were also administered the job knowledge and task proficiency tests. Cognitive ability scores (i.e., the AFQT) were available from the ASVAB. (The tests had been administered as the soldiers entered the Army.)
Analyses

To avoid the common method-variance problem associated with having the same rater providing the interpersonal factor and performance ratings, for each soldier ratee, one supervisor rater was randomly selected to provide the performance ratings, and the other supervisor rater then provided the interpersonal factor ratings. Similarly, for the peer ratings, one peer per ratee was randomly selected to contribute the performance ratings, and the other 1 to 4 peers (\(M = 2.10\)) then contributed the interpersonal factor ratings.

As mentioned earlier, causal models of job performance ratings were developed and tested separately for supervisor ratings and then peer ratings. Both causal models contained the following variables: general cognitive ability, technical proficiency, job knowledge, the interpersonal factors, and job performance ratings. As in the Borman et al. (1991) version of the Hunter (1983) model, ability was hypothesized to influence the performance ratings indirectly with a direct path to job knowledge, which in turn was hypothesized to have an indirect effect on the performance ratings with a direct path to technical proficiency. Technical proficiency was hypothesized to have a direct effect on the performance ratings. Finally, each of the composite interpersonal factors was hypothesized to have a direct effect on the performance ratings, with the Trustworthy/Dependable factor also hypothesized to have a direct effect on job knowledge. This last hypothesis was based on similar findings from the research by Borman et al. (1991).

To facilitate comparisons with previous studies, the revised Hunter (1983) model was first run for both rating models, and then the interpersonal factors were added. The fit of these various models, along with the variance accounted for in the job performance ratings, was evaluated and compared.

Specifically, the LISREL VII computer program (Jöreskog & Sörbom, 1989), which estimates parameters of hypothesized causal models, was used to analyze these data. To judge the fit of the path models to the data, chi-square, an adjusted goodness-of-fit index (AGFI), the root mean square error of approximation (RMSEA), and an expected cross-validation index (ECVI; Browne & Cudeck, 1993) were used for each model. The first two fit indexes indicate the likelihood that the hypothesized model could have produced the observed data. That is, they reflect the degree to which covariances estimated by the model reproduce the actual sample covariances among the observed variables. The RMSEA index is a discrepancy or residual measure per degree of freedom, and ECVI estimates the discrepancy between the final covariance matrix in the analyzed sample and the expected covariance matrix that would be obtained in another sample of the same size.

Observed correlations between the variables for the supervisor models appear above the diagonal in Table 2. These were derived by computing correlations for each of the five jobs separately and then creating meta-correlations across the five matrices. Each correlation for each job was transformed to a \(z\) score, and a weighted mean (i.e., by \(N-3\times\) the sample size) of these five \(z\) scores was calculated. Finally, the resulting \(z\) score was transformed back to a correlation to yield each of the meta-correlations in Table 2. The same procedure was followed for the peer models, and the corresponding uncorrected correlations are displayed above the diagonal in Table 3. For the LISREL analyses, we inputted these observed correlations, along with weighted means and weighted standard deviations (i.e., the pooled covariances) and the fixed error variances (i.e., \(1 - \text{reliabilities}\) \(\times\) variances). Reliability estimates were as follows: cognitive ability (.93), job knowledge (.87), task proficiency (.52), interpersonal factors and performance ratings (coefficients alpha as given in the Instruments section).

Rationale for the cognitive ability, job knowledge, and task proficiency reliability estimates can be found in Borman et al. (1991). For the interpersonal factor and performance ratings, we used coefficient alpha reliabilities primarily because their use leads to more conservative adjustments for error in latent variable models. There is little guidance regarding the estimation of error in latent variable models that contain ratings. Our belief is that interrater reliabilities likely provide an underestimate of reliability because different raters are typically not, strictly speaking, parallel. On the other hand, coefficients alpha probably overestimate “true” reliability because halo contributes to the magnitude of these estimates. Nonetheless, the coefficients alpha were used to provide the more conservative adjustments for error of the path coefficients.

Results

Baseline Hunter Models

The Borman et al. (1991) version of Hunter’s (1983) model was tested for both supervisor and peer ratings. Figure 1 depicts the resulting causal models, with the supervisor rating path coefficients appearing above and the peer coefficients below each path. Chi-square statistics for the two sets of data were as follows: \(\chi^2(3, N = 493) = 1.60\) (ns) for supervisor ratings and \(\chi^2(3, N = 631) = 4.35\) for peer ratings. Adjusted goodness-of-fit indices (AGFI) were .99 and .98. These results reflect a very good fit. The variance accounted for in the ratings by the models’ variables were 13% and 7%, respectively. The 13% for the supervisor model is fairly close to the 16% accounted for by the same variables in the supervi-

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ability</td>
<td></td>
<td>.44**</td>
<td>.14**</td>
<td>.05</td>
<td>.06</td>
<td>.03</td>
<td>-.04</td>
</tr>
<tr>
<td>Job Knowledge</td>
<td>.49</td>
<td></td>
<td>.36**</td>
<td>.18**</td>
<td>.15**</td>
<td>.05</td>
<td>.00</td>
</tr>
<tr>
<td>Technical proficiency</td>
<td>.20</td>
<td>.54</td>
<td></td>
<td>.24**</td>
<td>.07</td>
<td>.03</td>
<td>.07</td>
</tr>
<tr>
<td>Performance ratings</td>
<td>.05</td>
<td>.20</td>
<td>.35</td>
<td></td>
<td>.38**</td>
<td>.30**</td>
<td>-.20**</td>
</tr>
<tr>
<td>Dependable</td>
<td>.07</td>
<td>.18</td>
<td>.11</td>
<td>.43</td>
<td></td>
<td>.57**</td>
<td>-.42**</td>
</tr>
<tr>
<td>Friendly</td>
<td>.03</td>
<td>.06</td>
<td>.05</td>
<td>.35</td>
<td>.69</td>
<td></td>
<td>-.35**</td>
</tr>
<tr>
<td>Obnoxious</td>
<td>-.05</td>
<td>.00</td>
<td>.11</td>
<td>-.23</td>
<td>-.51</td>
<td>.43</td>
<td></td>
</tr>
</tbody>
</table>

Note: Entries are meta-correlations across five jobs. Uncorrected correlations appear above the diagonal, and corrected correlations are below the diagonal.

\* \(p < .05\). \** \(p < .01\).
Table 3
Variable Intercorrelations for Peer Models (N = 631)

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ability</td>
<td>—</td>
<td>.45**</td>
<td>.13**</td>
<td>.04</td>
<td>.10*</td>
<td>.09*</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2. Job knowledge</td>
<td>.50</td>
<td>—</td>
<td>.38**</td>
<td>.17**</td>
<td>.13**</td>
<td>.04</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3. Technical proficiency</td>
<td>.19</td>
<td>.56</td>
<td>—</td>
<td>.17**</td>
<td>.09*</td>
<td>.04</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>4. Performance ratings</td>
<td>.04</td>
<td>.19</td>
<td>.25</td>
<td>—</td>
<td>.27**</td>
<td>.22**</td>
<td>.22**</td>
<td>.00</td>
</tr>
<tr>
<td>5. Dependable</td>
<td>.11</td>
<td>.15</td>
<td>.14</td>
<td>.32</td>
<td>—</td>
<td>.66**</td>
<td>.43**</td>
<td>.35**</td>
</tr>
<tr>
<td>6. Friendly</td>
<td>.10</td>
<td>.05</td>
<td>.06</td>
<td>.26</td>
<td>.81</td>
<td>—</td>
<td>.47**</td>
<td>.35**</td>
</tr>
<tr>
<td>7. Obnoxious</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.26</td>
<td>.52</td>
<td>.58</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>8. Show-off</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.00</td>
<td>.05</td>
<td>.45</td>
<td>.73</td>
<td>—</td>
</tr>
</tbody>
</table>

*Note.* Entries are meta-correlations across five jobs. Uncorrected correlations appear above the diagonal, and corrected correlations appear below the diagonal.

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

isor model tested with the data from the research by Borman et al. The baseline peer model accounts for considerably less variance in the ratings, compared with the supervisor model.

**Expanded Models Including Interpersonal Factors**

Figure 2 contains results of a test of the expanded supervisor model that includes the three interpersonal factors: \( \chi^2(8, \ N = 493) = 10.7 \) (ns), AGFI = .97, RMSEA = .02, ECVI = .13. These results reflect a good fit. Variance accounted for in the ratings by the exogenous variables was 28%, more than twice the variance accounted for by the model without the interpersonal factors.

It is interesting that the largest path coefficient from the interpersonal variables to the performance ratings is for the Dependable/Trustworthy factor, virtually the same magnitude as the Technical Proficiency \( \rightarrow \) ratings path coefficient. The path from Friendly/Easy-To-Get-Along-With to the ratings and the Obnoxious/Complaining factor \( \rightarrow \) ratings path are substantially lower.

Figure 3 shows the corresponding expanded peer rating model path coefficients: \( \chi^2(10, \ N = 631) = 18.78 \) (p < .05), AGFI = .96, RMSEA = .05, ECVI = .17. This fit is not as good as with the supervisor rating model or either of the Hunter (1983) variable models. However, the AGFI and RMSEA represent at least a reasonably good fit. Variance accounted for in the ratings is 19%, lower than for the supervisor expanded model, but almost three times greater than when the interpersonal factors were not included in the peer model. The largest path from the interpersonal factors to the performance ratings is for the Dependable/Trustworthy, Obnoxious/Nasty, and Show-Off factors. The Friendly/Easy-To-Get-Along-With \( \rightarrow \) ratings path is near zero. The sign for the Show-off \( \rightarrow \) ratings path is counterintuitive and this variable appears to be acting as a suppressor variable. It is relatively highly correlated with Dependable and Obnoxious and has low correlations with the performance ratings.

In addition to the LISREL treatment of the data, we conducted regression analyses, with the performance ratings as the dependent variables and all other measures as independent variables. Results were very similar; the patterns of beta weights for both the supervisor and peer regression models were virtually identical to results emerging from the respective LISREL analyses.

**Discussion**

Hunter (1983) demonstrated that supervisory ratings appear to be substantially based on ratee technical

![Figure 1](image)

**Figure 1.** Revised Hunter causal models for supervisor and peer ratings. Path coefficients above the arrows are for the supervisor model, and path coefficients below the arrows are for the peer model. **p < .01 for the path coefficient.
performance (i.e., job knowledge and technical proficiency). His meta-analysis and our subsequent replications in Borman et al. (1991) and in the present study suggest that 13% to 17% of the variance in supervisor ratings can be accounted for by ratee ability, job knowledge, and technical proficiency. However, when variables of rater-ratee relationship and ratee personal characteristics are introduced into performance rating models, the

**Figure 2.** Expanded causal model including interpersonal factors for supervisor ratings. *p < .05 for the path coefficient. **p < .01 for the path coefficient.

**Figure 3.** Expanded causal model including interpersonal factors for peer ratings. *p < .05 for the path coefficient. **p < .01 for the path coefficient.
variance accounted for in the supervisor ratings increases by a factor of approximately 2 in Borman et al. and in this research. It is noteworthy that in each of the latter two studies different methods or different raters were used to measure the interpersonal variables and to collect the performance ratings. Further, in the peer model, variance accounted for in the ratings is lower but increases by a factor of more than 2½ when the interpersonal factors are added to the model.

In this study's supervisor-rating model, the largest direct paths to the performance ratings comes from the Dependable/Trustworthy and Technical Proficiency variables. Neither the Friendly/Easy-to-Get-Along-With nor the Obnoxious/Nasty variable shows a significant impact on supervisor performance ratings in the model. Results showing a substantial path from Dependable to the performance ratings are similar to the findings for Borman et al. (1991), where ratee Dependability measured in that research by a self-report temperament scale had a significant direct effect, and a substantial indirect effect through ratee problem behavior, on the supervisory performance ratings.

Results of the peer rating model are somewhat more difficult to interpret because of what appears to be a suppressor effect for one of the interpersonal variables. However, it is clear that the ratee Dependable/Trustworthy construct is an important contributor to performance ratings made by peers. Further, peers clearly rate their Obnoxious coworkers, characterized as being irritable, moody, mean, complaining, and rude, lower than they do their more pleasant peer counterparts.

As mentioned, the Show-Off variable may be acting as a suppressor in the peer model. The positive path from Show-Off to the performance ratings is suspect, because of this variable's high positive correlation with Obnoxious, fairly high negative correlation with Dependable, and near-zero correlation with the performance ratings. This finding is reminiscent of results presented by Collins and Schmidt (1993), who found cross-validated suppressor effects for several personality predictors of job performance. They suggest that the next step is to design research to discover the invalid variance components in other personality variables that are being partialed out by the suppressors. We echo that suggestion, but for the present research, it is simply very difficult to interpret the role of the Show-Off variable in the peer model.

For both the supervisor and peer models, the question arises, do the interpersonal variable–performance rating relationships indicate bias in the ratings? One way to address this question is to consider these variables of rater–ratee relationship and ratee characteristics not so much as biasing influences as what Borman and Motowidlo (1993) have called contextual performance factors. The Dependable/Trustworthy variable has elements of Borman and Motowidlo's third contextual area, Helping and Cooperating With Others, as well as aspects of their fourth area, Following Organizational Rules and Procedures. The Obnoxious/Nasty variable, prominent in the peer-rating model, corresponds with elements of the negative side of Helping and Cooperating With Others.

Similar to Organ (1988) in relation to organizational citizenship behavior and Brief and Motowidlo (1986) regarding prosocial organizational behavior, Borman and Motowidlo argued that contextual performance in organizations is important because it shapes "the organizational, social, and psychological context that serves as the critical catalyst for task activities and processes" (p. 71), and this kind of performance contributes importantly to organizational effectiveness. Seen in this light, the present results largely confirm the findings of MacKenzie, Podsakoff, and Fetter (1991) and Motowidlo and Van Scotter (1994). Both of these studies showed that supervisor raters weight contextual performance approximately as highly as task performance when making overall performance ratings.

Accordingly, there is evidence here that ratee contextual performance and task performance on the technical parts of these jobs both have a substantial impact on performance ratings. Also, for both the supervisor and peer-rating models, it is important to note that the Friendly/Easy-to-Get-Along-With variable, which has likability as one of its elements, has an insignificant direct path to the ratings (although the correlations between these two variables are .30 and .22, respectively). Overall, this study's results, particularly for the supervisor model, along with recent research concluding definitively that ratee race and sex have minimal effects on performance ratings (Oppler, Campbell, Pulakos, & Borman, 1992; Pulakos, White, Oppler, & Borman, 1989), shed a more favorable light on performance ratings in comparison with how they are usually characterized. Clearly, ratings contain error. Nonetheless, this study's findings and the other research cited are somewhat encouraging for the validity of performance evaluations, at least those ratings collected for research purposes.

Finally, comparisons between the supervisor and peer model results indicate both similarities and differences in the way these raters weight various factors. Both rater groups weight ratee dependability and task performance highly. However, ratee obnoxiousness appears to influence peer raters much more than it does supervisor raters. This discrepancy is likely due to differences in power and status in the relationships between peers versus between supervisors and subordinates. In general, peers might be exposed more often than supervisors to ratee behavior in this category. Differences between the supervisor and peer models do not support results of research by Tsui and O'hlott (1988) indicating highly sim-
ilar patterns of weights used by supervisor and peer raters. It should be noted, however, that the Tsui and Ohlott study did not examine the kinds of interpersonal factors included in the present research.

In sum, these and previous models of job performance ratings are instructive regarding the factors and cues raters use when making performance judgments, although it should be noted that these findings are derived from research-only field ratings. Performance appraisals for different purposes may result in different patterns of factors/cues being used. Nevertheless, Hunter, Schmidt, and colleagues have established that ratee ability, job knowledge, and technical proficiency do have a substantial influence on supervisory research ratings. Our work has shown that certain factors of rater-ratee relationships and ratee personal characteristics are also very important for supervisor and peer ratings, with the addition of those interpersonal factors increasing the variance accounted for in ratings by a factor of 2 to 2½. Significant similarities and differences in factor weights were evident across the supervisor and peer rating sources.

Future research on performance rating models should continue to examine the factors and cues raters use when making judgments about ratee performance. Research conducted in field settings, with interpersonal factors measured by methods different from those used to collect the performance ratings, will be especially helpful.

References
Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of


Pulakos, E. D. (1984). A comparison of rater training pro-


Received August 6, 1993
Revision received June 28, 1994
Accepted June 28, 1994