Evaluating the Performance of Paper People

Kevin R. Murphy
Colorado State University

Barbara M. Herr, Maura C. Lockhart, and Eammon Maguire
New York University

Several recent studies of performance appraisal have used the "paper people" approach, in which raters read performance vignettes and then rate the performance of several hypothetical ratees. This approach may lead to systematically different outcomes from studies in which ratings are based on the direct or indirect observation (e.g., via videotape) of behavior. One hundred and eleven studies published between 1975 and 1984 were grouped into five major substantive categories, and a comparative meta-analysis was used to contrast the outcomes of paper people studies to those of similar studies in which ratings were based on the observation of ratee behavior. Effect sizes were found to be significantly larger in paper people studies, although this difference was not uniform across all research areas. Results are discussed in terms of differences in signal-to-noise ratios across the two methods.

Many methods have been used in research on performance appraisal. One dimension for classifying these methods is based on whether they use a "paper people" approach, in which subjects read a series of vignettes containing scaled behavior descriptions and then rate the performance of one or more hypothetical ratees, or an approach that is based on the direct or indirect observation (i.e., via film or videotape) of the behavior of one or more ratees. The latter method has been used in studies ranging from those examining annual performance appraisals conducted for administrative purposes in organizations (e.g., Cleveland & Landy, 1981) to laboratory studies, in which the ratees are actors carrying out performance scripts (e.g., Murphy, Balzer, Kellam, & Armstrong, 1984). The critical difference between the two approaches is that in a paper people study, subjects receive written descriptions of the performance they are to evaluate, whereas in studies involving behavior observation, subjects must observe, interpret, and recall the behavior of ratees in the process of evaluating their performance.

The use of paper people in research on the interview has been strongly criticized (Gorman, Clover, & Doherty, 1978), although Dipboye, Stramler, and Fontenelle (1984) noted that this approach in fact simulates several features of the interview process. There are also grounds for criticizing the use of paper people in performance appraisal research. First, the question of whether raters observe or read summaries of the performance they are to evaluate has a substantial impact on the nature of the rating task. For example, raters' ability to accurately observe, encode, and recall performance is a critical concern in much of the recent research on performance appraisal (DeNisi, Meglino, & Cafferty, 1984; Ilgen & Feldman, 1983), but it is likely to be less relevant in evaluations of paper people then in evaluations of behavior one has actually observed. In particular, information one has read about another's behavior is likely to be represented differently in memory than information obtained by observing another's behavior (Tulving, 1983). It can be argued that research involving hypothetical ratees poses fundamentally different cognitive demands than research involving direct or indirect observation of performance.

Second, the information presented in a paper people experiment is generally less ambiguous than that obtained through behavior observation. The statement "John did X, Y, and Z, whereas Paul did only X" is less ambiguous, both with regard to the performance categories involved and to the level of performance, than the information obtained by observing an ongoing stream of behavior. Finally, the vignettes used in paper people research generally contain little irrelevant information; most of the information in these vignettes is directly relevant to the evaluation of the ratee's performance (Wendelken & Inn, 1981). In contrast, raters who observe ratee performance are typically confronted with a great deal of potentially irrelevant information (e.g., ratee attractiveness, behaviors not linked to performance); Banks and Murphy (1985) have argued that in many organizational settings, raters' ability to distinguish relevant from irrelevant information may be the most critical determinant of rating accuracy.

Although in theory, one might expect the paper people approach to lead to outcomes different from those of approaches involving direct or indirect observations of behavior, we know of little research comparing the outcomes of these two approaches to research on performance appraisal. The techniques of meta-analysis (Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982) are appropriate for addressing this question. The purpose of our study is to determine whether performance appraisal research using paper people leads to out-

The authors wish to thank several colleagues who provided prepublication drafts of papers for our review.

Correspondence concerning this article should be addressed to Kevin R. Murphy, Department of Psychology, Colorado State University, Fort Collins, Colorado 80523.
comes systematically different from those resulting when raters directly or indirectly observe performance. In particular, the present study sought to determine whether studies of paper people yield systematically larger or smaller effect sizes than similar studies involving behavior observation.

Design of the Study

The typical meta-analysis features a comprehensive review of the published and unpublished research in a particular domain. This study, which constitutes a comparative meta-analysis, focused on performance appraisal research for which a valid comparison of studies using performance vignettes versus observations of ratee behavior could be made. In particular, this review focused on areas within the extensive body of performance appraisal research for which comparable studies involving both approaches could be identified.

This review did not include research on interview judgments, on assessments of leader behavior (as distinct from job performance), or on performance in decision-making tasks or groups. To be included in our review, a study had to include real or simulated measures of job performance, and a comparison of different approaches, different groups, or rating outcomes under different conditions that could be used to estimate effect sizes. Thus, studies assessing the psychometric characteristics of a particular scale or method without explicitly comparing that method to some other (e.g., comparing a behaviorally anchored rating scale to a graphic scale) were not included in our analysis (see, for example, Goodale & Burke, 1975).

The first step in this study involved a review of research on performance appraisal published between 1975 and 1984 in the following journals: (a) Journal of Applied Psychology, (b) Organizational Behavior and Human Performance, (c) Personnel Psychology, (d) Psychological Bulletin, and (e) Academy of Management Journal. We also reviewed studies published prior to 1975 cited by more than one of the articles included in our 1975–1984 sample. This sample of older articles is not necessarily comprehensive or representative, but it contributes to the stability of our results by increasing the number of effect size estimates in our analysis. The resulting set of studies was sorted into five general categories:

1. Rater/Ratee characteristics consisted of research on the effects of personal characteristics of the rater or ratee (e.g., age, race, attractiveness, sex) on a variety of outcomes. Specific research topics included in this category were the effect of rater sex, effect of ratee sex, and effect of other rater or ratee characteristics on rating outcomes.

2. Rating scale format consisted of research on the effects of rating scale format on leniency, halo, and interrater agreement.

3. Training effects consisted of research on the effects of several types of rater training on the leniency and/or halo of ratings.

4. Performance level consisted of research in which level of ratee performance could be independently measured, manipulated, or controlled, and in which ratings were typically assessed as a function of known differences in performance levels.

5. Purpose of rating consisted of research on the effects of the apparent or actual purpose of rating on measures of leniency.

Hypotheses

Two hypotheses were offered regarding the size of the effects observed in studies involving paper people versus behavior observation.

1. Effect sizes will be systematically larger in paper people studies than in those involving behavior observation.

The rationale for this hypothesis is straightforward. Performance vignettes used in paper people studies typically present clearer, less ambiguous, and more relevant information; moreover, they appear to present a simpler task.

2. There will be an interaction between the study type and the content category.

This hypothesis implies that the difference between the outcomes of paper people versus behavior observation studies will not be uniform, but rather will vary according to the substantive area involved. For example, we may observe substantial differences in some areas (e.g., formats), but relatively small differences in others (e.g., training). No specific form was hypothesized for this interaction.

Method

Computation of Effect Size Estimates

To analyze the effects of research methods on study outcomes, these outcomes must be expressed using a single metric. The standardized mean difference, \( d \), was used as the common metric of effect size (Hunter et al., 1982). The effect size measure \( d \) represents the average group difference divided by the pooled standard deviation.

A large number of studies failed to report means and standard deviations. In several studies, formulas presented in Glass et al. (1981) and in Hunter et al. (1982) were used to re-express \( t \) values, \( r \) values, and \( F \) values in terms of \( d \). Nevertheless, it was often impossible to extract effect size estimates from studies (cf. Guzzo, Jackson, & Katzell, in press). In particular, studies reporting significance levels, but not reporting sufficient statistics needed for calculating \( d \), were dropped from our analysis. This category included studies that reported a statistically significant outcome, without reporting means, standard deviations, test statistics, or exact \( p \) values, as well as studies that merely reported that a particular test was not significant. Although methods exist for estimating \( d \) from significance tests, these estimates are known to be both biased and inexact (Glass et al., 1981).

In many studies, we could compute several effect size estimates, one for each dependent variable. Where several dependent variables were analyzed to address the same question, we pooled effect size estimates. Thus, our basic unit of analysis for each study examined was the content category into which the results in question could be classified. For example, Borman (1979) examined the effects of rating scale format on four separate measures of halo. Effect size estimates were calculated for each measure, and were pooled to arrive at a single effect size estimate. Borman and Dunnette (1975), on the other hand, examined the effects of rating scale format on six separate measures: (a) three measures of leniency, and (b) three measures of halo. Two separate pooled effect size measures, one describing leniency effects and one describing halo effects, were calculated for that study.

Final Sample

For the five content categories described earlier, we were able to identify 111 separate studies, which yielded 146 effect size estimates. Twenty-nine of these studies featured at least one experiment in which
Results

The effect sizes obtained from studies of effects of rater or ratee characteristics, rating scale format, training, performance level, and the purpose of rating are presented in Tables 1–5.

In general, the effects reported in these studies were not large; the overall mean effect size reported here was .34. Means and standard deviations for each of the major categories and subcategories are presented in Table 6. On the average, effect sizes the paper people approach was used; 41 of the effect size estimates analyzed here were obtained from these studies.
were larger in studies involving paper people than in those involving behavior observation (mean \( f \) values of 0.42 and 0.31, respectively), with substantial differences in effect size in research on effects of variations in performance level, and on the purpose of rating, and smaller differences elsewhere.

We used a fixed-effects analysis of variance (ANOVA) to test the significance of differences in effect size as a function of study type and content category. As can be seen in Table 7, the study type main effect, the content category main effect and the Type \( \times \) Category interaction were all significant. Study type (paper vs. behavior observation) accounted for a relatively small portion of the variance in effect sizes (\( \omega^2 = 0.02 \)); the content category main effect and the Type \( \times \) Category interaction accounted for more substantial portions of the variability in effect sizes (\( \omega^3 = 0.21 \) and 0.15, respectively).

### Discussion

As hypothesized, systematic differences were found in the outcomes of studies using the two research approaches (paper people vs. behavior observation). The average effect size was larger in paper people studies, but this difference was largely restricted to studies of the effects of variation in true performance level and of the effects of purpose of rating. Our results suggest that in some cases it does make a difference, even in terms of such a crude measure as the overall effect size, whether ratings are based on paper people or on behavior the rater has observed.

Our second hypothesis, that the difference between the outcomes of paper people studies and studies involving behavior observation would vary across content areas, was also supported. The Type \( \times \) Category interaction was both statistically significant and substantial (\( \omega^2 = 0.15 \)); tests of simple effects showed that effect sizes were significantly different in only one content category, purpose of rating. Even this result must be interpreted with caution, because there were only two paper people studies in this category. It should be noted, however, that there was little power for pairwise comparisons in any of the individual categories.

Our review suggests that the paper people approach is not widely used, except in assessing the effects of rater and ratee demographic characteristics, on ratings; 22 of the 41 effect sizes obtained from paper people studies were obtained from studies of the effects of rater or ratee characteristics. Contrary to the conclusions reached by Wendelken and Inn (1981), we found practically no difference in the effect sizes obtained in studies of rater and ratee characteristics involving paper people, as opposed to behavior observation (average effect sizes were .25 and .19, respectively).

The central difference between the two methods contrasted here can be described in terms analogous to signal and noise. In a typical paper people study, raters do not have to distinguish relevant from irrelevant information, do not have to interpret ambiguous information, do not have to retain information over long periods, and do not have to conduct performance appraisals concurrently with several other managerial tasks. In the field, the rater's main task may be to separate the signal (i.e., valid information about performance) from the noise. Even in laboratory studies in which subjects evaluate videotaped ratees, the subject's task may be considerably more complex than in a similarly designed paper people study. In addition, paper people studies may allow more powerful manipulations of experimental variables (stronger signal). For example, true performance levels can be manipulated to almost any degree in a paper people study; substantial differences in true performance levels may be much rarer in the field. In fact, given the differences between the tasks involved in the two approaches, it is surprising that the differences in effect size are not larger.

It is impossible to determine, on the basis of the available information, whether the presence of stronger effects in paper people studies is an indication of stronger signals (i.e., stronger manipulation of experimental variables) or less noise (i.e., less ambiguity in the rating task). Both explanations are plausible. For example, the significant difference in effect sizes in studies of the purpose of rating probably reflects less noise. The manipulation of purpose is generally similar across paper people and

### Table 4

<table>
<thead>
<tr>
<th>Study</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brief &amp; Wallace(1976)</td>
<td>1.37</td>
</tr>
<tr>
<td>Butterfield &amp; Powell(1981)</td>
<td>0.67</td>
</tr>
<tr>
<td>Cleveland &amp; Landy(1983)</td>
<td>0.65</td>
</tr>
<tr>
<td>DeNisi &amp; Stevens(1981)</td>
<td>1.44</td>
</tr>
<tr>
<td>Morrow, Larrenburg, Larson, Redfeather, &amp; Schoone (1983)</td>
<td>1.53</td>
</tr>
<tr>
<td>Rosen &amp; Jerdee(1974b)</td>
<td>0.27</td>
</tr>
<tr>
<td>Holzbach(1978)</td>
<td>0.21</td>
</tr>
<tr>
<td>Igel, Mitchell, &amp; Fredrickson (1981)</td>
<td>0.55</td>
</tr>
<tr>
<td>Leventhal, Perry, &amp; Abram(1977)</td>
<td>1.13</td>
</tr>
<tr>
<td>Mount (1984)</td>
<td>0.17</td>
</tr>
<tr>
<td>Schneier &amp; Beatty (1978)</td>
<td>0.04</td>
</tr>
<tr>
<td>Scott &amp; Hamer (1975)</td>
<td>0.35</td>
</tr>
<tr>
<td>Siegel(1982)</td>
<td>1.37</td>
</tr>
</tbody>
</table>

### Table 5

<table>
<thead>
<tr>
<th>Study</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeNisi, Cafferty, Williams, Blencoe, &amp; Meglino (1983)</td>
<td>2.10</td>
</tr>
<tr>
<td>Zedeck &amp; Cascio (1982)</td>
<td>1.37</td>
</tr>
<tr>
<td>Bernardin, Orban, &amp; Carlyle (1981)</td>
<td>0.82</td>
</tr>
<tr>
<td>Borresen (1967)</td>
<td>0.28</td>
</tr>
<tr>
<td>Centra (1976)</td>
<td>0.16</td>
</tr>
<tr>
<td>Hollander (1957)</td>
<td>0.14</td>
</tr>
<tr>
<td>Meier &amp; Feldhusen (1979)</td>
<td>0.15</td>
</tr>
<tr>
<td>Murphy, Balzer, Kellam, &amp; Armstrong (1984)</td>
<td>0.13</td>
</tr>
<tr>
<td>Sharon &amp; Bartlett (1969)</td>
<td>0.12</td>
</tr>
</tbody>
</table>
behavior observation studies (e.g., ratings are done either for research or for the purpose of making administrative decisions), but the credibility of that manipulation may be higher in laboratory studies than in real organizations. On the other hand, manipulations of level of performance are probably more extreme (more signal) in paper people studies than in behavior observation studies involving real performance appraisals; there were not enough studies of each type to test this hypothesis adequately. Nevertheless, the results obtained here suggest that systematically different results will be obtained, at least with regard to the magnitude of the effects, when evaluating the performance of hypothetical versus real ratees. In other words, the paper people approach may lead to an overestimate of the effects of some variables (e.g., purpose of rating) on the evaluation of performance.

References


Bernardin, H. J. (1981, August). Perfor-


and sex as determinants of ratings by potential employees in a simulated work shaping task. *Journal of Applied Psychology*, 59, 705–711.


---

**Integrating Personality and Social Psychology: Call for Papers**

The editors of the *Journal of Personality and Social Psychology* intend to publish a special issue devoted to papers demonstrating that social behavior is best understood by integrating the diverse concerns of the three sections of *JPSP.* Papers are invited that deal with particular substantive issues crossing the boundaries of individual differences, social cognition, and interpersonal relations. Papers should represent practical demonstrations that the diverse concerns of this journal belong together in a full understanding of social behavior. We seek previously unpublished contributions, primarily empirical studies, but we are also amenable to syntheses of long-term research programs and to innovative theoretical statements.

Contributions intended for the special issue should be sent to the guest editor:

John F. Kihlstrom, PhD  
W. J. Brogden Psychology Building  
University of Wisconsin  
1202 West Johnson Street  
Madison, Wisconsin 53706

Interested authors should send an abstract of their article to the guest editor by September 1, 1986 and plan to submit a completed manuscript by December 1, 1986, at which time the paper will become subject to the usual peer review process.

*Journal of Personality and Social Psychology* editors:

- **Attitudes and Social Cognition**  
  - Steven J. Sherman  
  - Charles M. Judd

- **Interpersonal Relations and Group Processes**  
  - Harry T. Reis  
  - Norbert L. Kerr

- **Personality Processes and Individual Differences**  
  - Irwin G. Sarason  
  - Edward F. Diener  
  - Warren H. Jones