A META-ANALYSIS OF SELF–SUPERVISOR, SELF–PEER, AND PEER–SUPERVISOR RATINGS

MICHAEL M. HARRIS, JOHN SCHAUBROECK
Krannert Graduate School of Management
Purdue University

Reviews of self–supervisor, self–peer, and peer–supervisor ratings have generally concluded that there is at best a modest correlation between different rating sources. Nevertheless, there has been much inconsistency across studies. Accordingly, a meta-analysis was conducted. The results indicated a relatively high correlation between peer and supervisor ratings ($\rho = .62$) but only a moderate correlation between self–supervisor ($\rho = .35$) and self–peer ratings ($\rho = .36$). While rating format (dimensional versus global) and rating scale (trait versus behavioral) had little impact as moderators, job type (managerial/professional versus blue-collar/service) did seem to moderate self–peer and self–supervisor ratings.

The use of multiple sources for performance ratings has gained considerable acceptance in the past two decades. Numerous advantages of using multiple raters have been cited: for example, enhanced ability to observe and measure various job facets (Borman, 1974; Henderson, 1984), greater reliability, fairness, and ratee acceptance (Latham & Wexley, 1982), and improved defensibility of the performance appraisal program from a legal standpoint (Bernardin & Beatty, 1984). The literature on multiple raters has focused on self–ratings. A number of scholars have argued that self–ratings can promote personal development, improve communication between supervisors and subordinates, and clarify differences of opinion between supervisors and other managers (Carroll & Schneier, 1982; Cummings & Schwab, 1973).

Despite the alleged gains from self–ratings, empirical research shows frequent lack of agreement between self–ratings and those provided by other sources. Mabe and West (1982) reviewed a number of studies and found, on average, a low correlation between self–ratings and others' ratings, including supervisor and peer appraisals. Thornton (1980), in his summary of self–ratings, concluded that "individuals have a significantly different view of their own job performance than that held by other people" (p. 268). Nevertheless, both Mabe and West and Thornton found that research often

---

The authors were equal contributors to this manuscript; order of listing was determined alphabetically.

The authors would like to thank Michael A. Campion for his helpful comments.

Correspondence and requests for reprints should be addressed to Michael M. Harris, Krannert Graduate School of Management, Purdue University, West Lafayette, IN 47907.

COPYRIGHT © 1988 PERSONNEL PSYCHOLOGY, INC.
led to conflicting results. Indeed, a casual inspection of research on self-supervisor ratings shows a wide range of findings. For example, Williams and Seiler (1973) reported an average self-supervisor correlation of .60; Pym and Auld (1965) found an average self-supervisor correlation of .56 across three independent studies. Conversely, Klimoski and London (1974) reported an average self-supervisor correlation of .05, and Ferris, Yates, Gilmore, and Rowland (1985) found a self-supervisor correlation of .02. Others (e.g., Heneman, 1974) have found moderate correlations between self-supervisor ratings. Thus, no firm conclusions can be drawn regarding the extent of self-supervisor agreement.

Low correlations between different raters are not unique to comparisons with self-ratings. On the one hand, some researchers have observed relatively high peer-supervisor correlations: for example, Lawler (1967) reported an average correlation of .57. On the other hand, others have found much lower correlations between peer and supervisor performance ratings: Springer (1953), for example, reported an average correlation of .28, and Tucker, Cline, and Schmitt (1967) reported .21. Thus, Landy and Farr (1980) have concluded that different raters exhibit moderate to low agreement. Moreover, they have suggested that the limited correspondence between self-ratings and those from other sources is not due to any unique differences beyond that to be expected for multiple raters in general.

It is important to note that a number of scholars have asserted that the general lack of convergence between different raters is neither surprising nor problematic. Supporters of this position argue that different raters may observe different dimensions of performance, or have different definitions of effective performance and, thus, arrive at different assessments of the same individual’s performance (Borman, 1974; Landy & Farr, 1980). As Wexley and Klimoski (1984) pointed out, however, a definitive test of this hypothesis has not been conducted.

A review of the literature suggests that a number of explanations have been offered as to why low correlations exist between different raters. These can be sorted into at least three categories: (1) egocentric bias, (2) differences in organizational level, (3) observational opportunities. What follows is a brief discussion of each of these explanations.

**Egocentric Bias**

While several different explanations fall under this rubric, the underlying premise is that ratee (or self-) ratings are biased in some fashion, while other raters (e.g., peers and supervisor) share a set of common perceptions. One of the most commonly discussed versions of egocentric bias is the main effect of defensiveness, wherein a self-rater is inclined to inflate his/her rating in order to enhance the evaluation (Holzbach, 1978;
Steel & Ovalle, 1984). This would lead the self-ratings to have a restricted range, thereby attenuating the correlation between self- and others’ ratings. Correlations between others’ ratings (e.g., peers and supervisors) would be considerably higher as there would be no range restriction. However, correcting for range restriction should eliminate any differences between correlations based on self- and others’ ratings and the correlations based on different observers’ ratings.

A second version of the egocentric bias explanation assumes that ratee defensiveness may be moderated by other variables, such as self-esteem (Baird, 1977; Kay, Meyer, & French, 1965). If the impact of ratee defensiveness depends on a third variable, such as self-esteem, the order of self-ratings may be affected. That is, high self-esteem ratees may inflate their ratings, while low self-esteem ratees may not. If the third variable is not taken into account, correcting for range restriction will not completely eradicate the egocentric bias, and correlations between observer ratings would remain higher than correlations between self- and others’ ratings.

A third type of egocentric bias is posited by attribution theory (DeVader, Bateson, & Lord, 1986; Jones & Nisbett, 1972). According to this theory, actors (i.e., self-raters) attribute good performance to their own behavior and poor performance to environmental factors. Conversely, observers (e.g., peers and supervisors) attribute good performance to environmental factors and poor performance to the actors’ dispositions. Hence, this theory would suggest that self-ratings will correlate poorly with ratings provided by other sources. Correcting for range restriction should have little effect. However, there should be substantial agreement between observers (i.e., peers and supervisors).

In addition to predictions about the correlations, all three versions of the egocentric bias explanation would predict differences in mean ratings. Specifically, self-ratings should be significantly higher than either peer or supervisor ratings.

Differences in Organizational Level

Two versions of the organizational-level explanation for rater disagreement exist. Some researchers (e.g., Klimoski & London, 1974; Zammuto, London & Rowland, 1982) have asserted that raters at different levels weight performance dimensions differently. This suggests that raters at different levels (e.g., peers and supervisors) disagree on the overall rating, but they presumably would agree on dimensional ratings. A second version of this explanation maintains that raters at different levels define and measure performance differently (e.g., Borman, 1974; Landy, Farr, Saal, & Freytag, 1976). This latter hypothesis would suggest that raters at different levels disagree on both dimensions and overall ratings. At the
same time, both explanations suggest that raters on the same level (e.g., self and peers) would provide similar ratings. Note that neither version of the organizational-level explanation would necessarily suggest differences in the mean ratings of self and others.

Observational Opportunities

A third explanation for lack of agreement across different raters concerns differing observational opportunities. Specifically, peers are hypothesized to have more opportunities to observe ratees and at more revealing times than do supervisors (Latham & Wexley, 1982; Lawler, 1967). This explanation implies that supervisors disagree with a ratee because they have few opportunities to observe the individual's performance, and the behaviors they do observe do not reflect true performance levels. Thus, one would expect self-supervisor and peer-supervisor ratings to exhibit far lower correlations than peer-self ratings. However, as with the organizational-level explanation, no differences between mean self-, peer, and supervisor ratings would be expected.

While the three basic explanations described above have been presented as mutually exclusive, it is possible that several are operating simultaneously. For example, Latham and Wexley (1982) provide an observational opportu-
nities explanation for the lack of agreement between peers and supervisors and suggest that defensiveness may be the cause of peer-self disagreement. Accordingly, while a direct test of these alternative hypotheses is not possible without measurement of proposed moderators, comparison of average correlations for different rater pairs can provide some direction and insight into which explanation may account for low correlations between raters.

The purpose of this paper was to conduct a quantitative review of empirical research on rater agreement. While a number of qualitative reviews have been reported (e.g., Landy & Farr, 1980; Thornton, 1980), a definitive estimate of rater agreement has not been ascertained.\(^1\) Furthermore, the apparent lack of consistent findings regarding rater agreement may be explained by sampling error, as many of these studies are based on small sample sizes. For example, Pym and Auld (1965) had an average of 70 subjects in each sample; Ferris et al. (1985) reported an \(N\) of 81. Thus, one explanation for the lack of consistency from study to study is that sampling error is affecting the results.

A second explanation for the apparent lack of consistency between studies is the nature of the job. That is, blue-collar and service jobs are

\(^1\)The one quantitative review (Mabe & West, 1982) did not separate out self-peer and self-supervisor correlations, nor did it examine peer-supervisor correlations.
relatively routine, and performance is well defined compared with managerial or professional jobs (Campbell, Dunnette, Lawler, & Weick, 1970; Cascio, 1987). Hence, it is to be expected that there will be a higher correlation between different raters for blue-collar/service jobs than for managerial/professional jobs.

Two other moderators are predicted by the explanations discussed above. As described by the first organizational-level hypothesis, different raters may weight dimensions differently. If this is the case, one would expect greater agreement on dimensional ratings than on global ratings. Hence, another potential moderator is rating format (dimensional versus global). Second, rating scale may be an important moderator whereby rater agreement will be higher for behavioral than for trait scales. This follows from the second version of the organizational-level hypothesis, which asserts that different raters may have different definitions of effective job performance (Borman, 1974; Zedeck & Baker, 1972). Hence, the more well defined and behaviorally based the rating scale, the greater the agreement expected between different raters.

In summary, the present study used meta-analysis to review the literature regarding the correlation between self–supervisor, self–peer, and peer–supervisor ratings. Specifically, the questions of interest were as follows:

1. (a) Overall, what is the average correlation between self–supervisor, self–peer, and peer–supervisor ratings? (b) What is the average difference between supervisor, peer, and self-ratings? (c) Will the data support an egocentric-bias, organizational-level, or observational-opportunities explanation for rater disagreement?
2. Can differences between studies within each rater pair be attributed to statistical artifacts such as sampling error?
3. Assuming that differences between studies cannot be explained by statistical artifacts, can the variance be explained by moderators such as job type (managerial/professional versus blue-collar/service), rating format (dimensional versus global), and rating scale (behavioral versus trait)?

Method

Analysis

The Hunter, Schmidt, and Jackson (1982) meta-analytic procedure designed to correct for unreliability and range restriction with sporadic information was employed for the correlations. This procedure involves a sequence of computations. In step 1, the sample weighted average correlation and the sample weighted variance of the correlations is calculated.
Next, the expected variance due to sampling error is determined, and this is subtracted from the observed variance.

Step 2 of the meta-analysis involves estimating reliability and range restriction across the variables. Once this is done, the average correlation is corrected for unreliability and range restriction, as is the variance. The formulas for these estimates may be found in Hunter et al. (1982). In accord with the main effect defensiveness hypothesis, self–other correlations were corrected for both of these artifacts. However, since there was no theoretical reason to correct for range restriction for peer–supervisor correlations, only corrections for measurement error were performed for this category.

The third step examines whether variation between studies is due to artifacts or is caused by moderator variables. The 25% rule was used, whereby a moderator search was initiated if the ratio of unexplained variance (i.e., variance remaining after correcting for sampling error, measurement error, and range restriction) to total variance was greater than 25%.

The fourth step involves testing for moderators. Towards this end, effect sizes are subgrouped and a separate analysis is performed on each subgroup. Supportive evidence for the hypothesized moderator variable exists when (1) the mean correlation differs from subgroup to subgroup and (2) the corrected variance is lower on average in the subgroups than for the overall set.

Mean differences between peer, supervisor, and self-ratings were converted into a $d$ statistic. Procedures described by Hunter et al. (1982) were then used to obtain the average effect size and true variance. The $d$ statistic was calculated such that the first rater in each rater pair (e.g., self in self–supervisor ratings) was associated with the higher ratings.

Sample


---

2 Note that the source of the range restriction differs from the traditional situation wherein range restriction is due to a lack of complete data. For the present situation, we computed a ratio of the standard deviation of self-ratings to the standard deviation of others’ ratings (either peers or supervisors). This ratio is referred to as $\tau$ in Hunter et al. (1982) and is used along with other formulas to correct for range restriction.

3 Because Hunter et al. (1982) do not provide formulas for correcting for measurement error with sporadic information, we used the average reliabilities listed in Table 1.
**Journal of Vocational Behavior** (1970–1986), *Organizational Behavior and Human Decision Processes* (Formerly Organizational Behavior and Human Performance) (1966–1986), and *Personnel Psychology* (1956–1986). Conference papers from the Midwest and National proceedings of the Academy of Management were also examined. Computer-based searches were conducted on Psyc-Info (Psychological Abstracts) for the period since 1967 and the ABI Inform data base for the period since 1975. References for other studies were obtained from these sources to expand the research base.

Three judgments were made concerning the collection and analysis of data. First, only those reliability estimates based on accepted formulas were selected to determine the mean reliabilities across studies. This excluded, among other estimates, unadjusted average inter-item correlation coefficients. Second, laboratory studies were deliberately excluded from the analysis. This was done because the conditions under which laboratory ratings occur grossly exaggerate the observation time factor. More importantly, assigned roles cannot be taken to reflect true perspectives of supervisors and peers (Ilgen & Favero, 1985). Finally, some studies contained multiple measures of performance. Since these measures were typically not independent, the effect sizes were averaged and the mean effect size was used in the meta-analysis.

**Results**

The literature search produced a total of 36 independent self–supervisor correlations, 23 independent peer–supervisor correlations, and 11 independent self–peer correlations (approximately 5 studies did not provide either means or correlations). Table 1 shows the average correlations, reliabilities, and range restriction for the three rater combinations. The reliabilities are based on internal consistency (alpha) measures. It is noteworthy that the average reliabilities are substantially higher than the .60 figure used by Schmidt and his colleagues (e.g., see Pearlman, Schmidt, & Hunter, 1980). There are three reasons for using the more conservative figures found here. First, Schmidt and his colleagues based their reliability estimates on the correlations between ratings from different supervisors at different points in time. In the present study, only internal consistency measures were found. Second, the .60 figure used by other researchers assumes that one source of error is the use of different raters. In the present case, at least for self-ratings, the rater did not change. Hence, a reliability of .60 would be an underestimate. Finally, there has been considerable debate over the accuracy of .60 as an estimate of reliability (Sackett, Tenopyr, Schmitt, & Kehoe, 1985; Schmitt & Schneider, 1983). Use of the higher, and hence more conservative, figures therefore seemed in order.
<table>
<thead>
<tr>
<th>Rater combination</th>
<th>Total sample size</th>
<th>Number of correlations</th>
<th>$a^a$</th>
<th>$b$</th>
<th>$u$</th>
<th>Mean of $r$ correlation</th>
<th>Standard deviations</th>
<th>Lower bound</th>
<th>Upper bound</th>
<th>% Unexplained variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-supervisor</td>
<td>3,957</td>
<td>36</td>
<td>.90(7)</td>
<td>.91(7)</td>
<td>.78(8)</td>
<td>.35(.22)</td>
<td>.11(.12)</td>
<td>.17</td>
<td>.53</td>
<td>33</td>
</tr>
<tr>
<td>Peer-supervisor*</td>
<td>2,643</td>
<td>23</td>
<td>.87(5)</td>
<td>.89(5)</td>
<td>NA</td>
<td>.62(.48)</td>
<td>.24(.22)</td>
<td>.22</td>
<td>1.00</td>
<td>86</td>
</tr>
<tr>
<td>Self-Peer</td>
<td>984</td>
<td>11</td>
<td>.90(7)</td>
<td>.87(5)</td>
<td>.84(3)</td>
<td>.36(.24)</td>
<td>.19(.13)</td>
<td>.05</td>
<td>.67</td>
<td>82</td>
</tr>
</tbody>
</table>

*a, b, and $u$ represent, respectively, average reliability for the first rater in each combination, the second rater in each combination, and the ratio between the standard deviation of the first rater and the standard deviation of the second rater in each pair (see Hunter et al., 1982, for further explanation). Figures in parentheses indicate number of studies involved in calculations.

*Numbers in parentheses include corrections for sampling error; numbers not in parentheses corrected for all possible artifacts.

Means and variance corrected for measurement error only; see text for explanation.
Because there was only one self–peer study reporting reliabilities, estimates of peer–rating reliability were obtained from the peer–supervisor studies; likewise, estimates of self–rating reliability were taken from self–supervisor studies.

As shown in Table 1, all three rater combinations had lower 90% confidence intervals that were greater than zero. Hence, the correlation between ratings from different raters is statistically significant. However, there were marked differences between rater combinations: The peer–supervisor correlation ($\rho = .62$) was substantially higher than either the self–supervisor ($\rho = .35$) or the self–peer correlation ($\rho = .36$). Finally, all three rater combinations contained more than 25% unexplained variance even after sampling error, measurement error, and range restriction were taken into account; accordingly, moderator variables were examined.

Table 2 summarizes the moderator analysis for self–supervisor correlations. Consistent with the hypothesis, there was slightly more agreement for dimensional ($\rho = .36$) than for global ($\rho = .29$) ratings. However, there was no decrease in variance across the subgroups. In accord with the rating-scale hypothesis, there was also a small difference between trait and behavioral scales, wherein greater agreement occurred for the latter. While the ratio of unexplained to total variance was less than 25% (15%) for behavioral ratings, there was no decrease in sample-size weighted variance averaged across subgroups compared with the whole set. Thus, rating scale and rating format were not significant moderators for self–supervisor correlations. However, job type appeared to be a significant moderator. When sampling error and other artifacts were taken into account, no variance remained for managerial/professional employees. Although substantial variance remained for blue-collar/service workers, the average corrected variance across the subgroups was lower than for the set as a whole. Moreover, there was lower agreement for managerial/professional employees ($\rho = .27$) than for blue-collar/service employees ($\rho = .42$). Hence, job type did seem to moderate self–supervisor correlations.

Table 3 summarizes the results of the moderator analysis for peer–supervisor correlations. Contrary to expectations, slightly higher agreement was observed for global than for dimensional ratings; nonetheless, there was no decrease in variance across subsets. More surprising was the finding that greater agreement was obtained for trait ($\rho = .64$) than for behavioral ratings ($\rho = .53$). There was only a slight reduction in the sample-size weighted average variance across subsets as compared with the variance in the whole set ($\bar{\rho} = .23$ across subsets as compared with .24 for the whole set). Finally, job type did not seem to affect the amount of agreement.

Due to the small number of studies (2) using either behaviorally based rating scales or dimensional formats, these variables were not investigated
<table>
<thead>
<tr>
<th>Moderator</th>
<th>Total sample size</th>
<th>Number of correlations</th>
<th>Mean* correlation</th>
<th>Standard deviation</th>
<th>90% Confidence intervals</th>
<th>% Unexplained variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating format†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimensional Global</td>
<td>2,542</td>
<td>29</td>
<td>.36(.23)</td>
<td>.16(.15)</td>
<td>.10 .62</td>
<td>50</td>
</tr>
<tr>
<td>Rating scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trait</td>
<td>2,899</td>
<td>21</td>
<td>.32(.21)</td>
<td>.12(.12)</td>
<td>.12 .52</td>
<td>46</td>
</tr>
<tr>
<td>Behavioral</td>
<td>860</td>
<td>13</td>
<td>.43(.28)</td>
<td>.08(.13)</td>
<td>.30 .56</td>
<td>15</td>
</tr>
<tr>
<td>Job type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Managerial/professional</td>
<td>1,935</td>
<td>11</td>
<td>.27(.17)</td>
<td>.00(.04)</td>
<td>.27 .27</td>
<td>00</td>
</tr>
<tr>
<td>Blue-collar/service</td>
<td>2,022</td>
<td>25</td>
<td>.42(.27)</td>
<td>.16(.15)</td>
<td>.16 .68</td>
<td>45</td>
</tr>
</tbody>
</table>

*Numbers in parentheses include corrections for sampling error; numbers not in parentheses corrected for all possible artifacts.
†Combined sample size larger than total sample since both dimensional and global ratings are included from same studies.
### Table 3
**Moderator Analysis Results: Peer–Supervisor Correlations**

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Total sample size</th>
<th>Number of correlations</th>
<th>Mean* correlation</th>
<th>Standard deviation</th>
<th>90% Confidence intervals</th>
<th>Unexplained variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating format</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimensional</td>
<td>1,331</td>
<td>14</td>
<td>.57 (.44)</td>
<td>.23 (20)</td>
<td>.19 .95</td>
<td>80</td>
</tr>
<tr>
<td>Global</td>
<td>1,914</td>
<td>15</td>
<td>.65 (.50)</td>
<td>.25 (21)</td>
<td>.24 1.00</td>
<td>86</td>
</tr>
<tr>
<td>Rating scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trait</td>
<td>2,137</td>
<td>18</td>
<td>.64 (.49)</td>
<td>.24 (23)</td>
<td>.24 1.00</td>
<td>88</td>
</tr>
<tr>
<td>Behavioral</td>
<td>506</td>
<td>5</td>
<td>.53 (.41)</td>
<td>.20 (18)</td>
<td>.20 .86</td>
<td>80</td>
</tr>
<tr>
<td>Job type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Managerial/professional</td>
<td>905</td>
<td>8</td>
<td>.64 (.48)</td>
<td>.24 (21)</td>
<td>.24 1.00</td>
<td>80</td>
</tr>
<tr>
<td>Blue-collar/service</td>
<td>1,738</td>
<td>15</td>
<td>.62 (.47)</td>
<td>.27 (23)</td>
<td>.17 1.00</td>
<td>87</td>
</tr>
</tbody>
</table>

*Numbers in parentheses include corrections for sampling error; numbers not in parentheses corrected for all possible artifacts.

### Table 4
**Moderator Analysis Results: Self–Peer Correlations**

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Total sample size</th>
<th>Number of correlations</th>
<th>Mean* correlation</th>
<th>Standard deviation</th>
<th>90% Confidence intervals</th>
<th>Unexplained variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Managerial/professional</td>
<td>534</td>
<td>6</td>
<td>.31 (.20)</td>
<td>.00 (.00)</td>
<td>.31 .31</td>
<td>00</td>
</tr>
<tr>
<td>Blue-collar/service</td>
<td>380</td>
<td>5</td>
<td>.40 (.26)</td>
<td>.28 (.19)</td>
<td>-.06 .86</td>
<td>14</td>
</tr>
</tbody>
</table>

*There were too few studies to test either rating format or rating scale as moderators.

*Numbers in parentheses include corrections for sampling error; numbers not in parentheses corrected for all possible artifacts.
as moderators for self–peer correlations. Thus, only job type was considered; the results are shown in Table 4.

As in the self–supervisor category, little or no variance remained for either subgroup of self–peer correlations when sampling error and other artifacts were taken into account. Accordingly, there was a reduction in average variance across subsets as compared with the whole set, while the average correlation was higher for blue-collar/service workers ($\rho = .40$) than for managerial/professional employees ($\rho = .31$). Thus, job type did meet the criteria for acceptance as a moderator.

A total of 18 independent self–supervisor means, 7 independent peer–supervisor means, and 4 independent self–peer means were located. As predicted by the egocentric-bias explanation, there was a fairly large difference between self- and supervisor ratings ($d = .70$). However, the true variance was also large ($\sigma^2 = .25$). Thus, the difference in means between self- and supervisor ratings was not statistically significant. Similarly, while on average self-ratings were somewhat higher than peer ratings ($d = .28$), a substantial amount of variance remained even after correcting for sampling error ($\sigma^2 = .09$). Hence, the difference in mean self–peer ratings was nonsignificant. Finally, while peers provided somewhat higher ratings than did supervisors ($d = .23$), the difference was not significant, given the large variance ($\sigma^2 = .11$).  

**Discussion**

The first question of interest concerned the average correlation and mean difference between rater pairs. Overall, the data suggest that after correcting for measurement error and range restriction, moderate agreement between self–peer ($\rho = .36$) and self–supervisor ($\rho = .35$) ratings exists. There was much higher agreement between peers and supervisors ($\rho = .62$). Nonetheless, the lower 90% confidence interval was greater than zero for all three rater combinations. The results of this study then are somewhat contrary to conclusions by Landy and Farr (1980). As noted earlier, they concluded that different raters exhibit low to moderate agreement. The present results indicate that while this is true for self-ratings compared with ratings by others, peer–supervisor ratings demonstrate a far higher average correlation. This discrepancy between a meta-analytic-based literature review and a traditional narrative review is not surprising, given the differences between the two procedures. The present results indicate the value of conducting quantitative reviews.

---

4 A moderator analysis subgrouping by job type for self–supervisor means revealed no reduction in variance.
In terms of mean differences, on average, self-ratings were over a half standard deviation higher than supervisor ratings and approximately one-quarter of a standard deviation higher than peer ratings. Nevertheless, given the large variances, the lower-bound confidence intervals included zero in both cases. However, examination of the effect sizes showed that in only one case were the self-ratings lower than supervisor ratings.

The second question concerned whether or not differences between studies within each rater pair could be attributed to statistical artifacts. The results indicated that in all cases substantial amounts of variance still remained, even when the appropriate corrections were made. Accordingly, moderator analyses were conducted.

Of the three moderators, only job type showed any meaningful effects. Specifically, self–supervisor and self–peer correlations were lower for managerial/professional employees than for blue-collar/service employees, and no true variance existed for the former category. However, such effects were not obtained for peer–supervisor correlations. This suggests that while incumbents of managerial/professional jobs have very different views of their performance than do others, observers exhibit far more agreement. One possible explanation is that egocentric bias is more likely to occur in ambiguous contexts (i.e., managerial/professional jobs) than in well-defined jobs (i.e., blue-collar/service). Whatever the explanation, job type does not appear to affect observer agreement; further research is needed to test why this is the case. The one instance where rating scale met the criteria for a moderator should be viewed cautiously for several reasons. First, only a small number of studies (5) using behaviorally based rating scales were available, and the total sample size was small ($n = 506$). Second, the decrease in amount of variance across subsets was very small. Thus, rating scale was at best a relatively minor moderator of peer–supervisor agreement. In answer to the third question then, moderators did help account for some of the variance.

Recall earlier that three basic explanations were reviewed as to why disagreement between raters may occur. The present findings are consistent with two versions of the egocentric-bias theory. Specifically, in accord with attribution theory (DeVader et al., 1986; Jones & Nisbett, 1972), observers (i.e., peer–supervisor combinations) displayed higher agreement than did "actors" with observers (i.e., self–supervisor and self–peer combinations). The findings also mesh with the moderated defensiveness explanation: even after the range restriction corrections, observer-observer ratings (i.e., peer–supervisor combinations) demonstrated much higher agreement than self-observer ratings.

Conversely, there was no support for the main-effect-of-defensiveness explanation as correcting for range restriction did not reduce the difference
between self-peer/self-supervisor correlations and peer-supervisor correlations. Nor was there evidence that observational differences accounted for rater disagreement as there was a relatively high correlation between peer-supervisor ratings. Finally, there was no support for the organizational-level hypotheses: an almost identical correlation was found between ratings by individuals at the same level (self-raters and peers) and individuals at different levels (self and supervisors). Moreover, the highest correlation was found for raters at different levels (peer-supervisor). Nevertheless, further research is needed to conduct more direct tests for all possible explanations of rater disagreement.

It is also noteworthy that, with one minor exception, neither rating format nor rating scale were significant moderators. Other meta-analyses of performance ratings have come to similar conclusions. For example, Kraiger and Ford (1985) found that rating scale did not moderate ratee race effects; Heneman (1986) also reported a lack of support for rating format as a moderator. Hence, despite what would seem to be critical moderators, research has found rating instrumentation to be less useful in identifying sources of variation than was previously thought to be the case. Conversely, the findings regarding job type suggest that this may be a moderator of importance that other researchers in performance appraisal should examine.

Although there are a number of potential reasons why a substantial amount of variance between studies remained in most instances, two seem most plausible. First, it seems likely that there are differences between studies in regard to criterion contamination or rater independence. Specifically, studies may have differed as to how much raters knew about others’ ratings. In some studies, subtle forms of criterion contamination may have existed. For example, it is possible that a few studies involving self-ratings were conducted shortly after performance appraisal reviews, wherein ratees had substantial information regarding supervisor ratings. In other instances, employees may have received informal feedback, while in yet other studies, employees may have had little or no feedback of any sort. Unfortunately, very few studies report enough detail about the amount or type of feedback to test whether this variable moderates rater agreement.

A second explanation for variance between studies may be differences in raters’ opportunity to observe job performance. For example, workers may have more interaction with peers in some cases than in others, thereby leading to differential agreement with supervisors. Jobs with much task interdependence may lead to greater agreement between self- and peer ratings than jobs where workers perform independently.

Other potential moderators include such things as purpose of rating (e.g., Zedeck & Cascio, 1982), amount of rater training (e.g., Smith, 1986), and rater motivation (Bernardin & Beatty, 1984). However, a review of
our studies showed that these variables were also rarely, if ever, reported. Hence, no moderator analyses could be performed. Moreover, McIntyre, Smith, and Hassett (1984) found that purpose of rating had little or no effect on ratings, and many studies have failed to show that training increases the accuracy of performance ratings, particularly when only a lecture approach is used (Smith, 1986). Further theory development and empirical research will be necessary to achieve greater knowledge about which conditions affect rater agreement.

Although one can always argue that a “file-drawer” problem may exist such than inclusion of missing studies would change the results, this is unlikely here. First, an extensive literature review was conducted. Second, Rosenthal’s (1979) index suggests that it would take many studies to dramatically alter the results, given the size of the correlations and the total sample size found here.

Future research needs to address two issues. First, a more direct test of these competing hypotheses is needed. That is, a study should be done wherein different raters (e.g., peers, supervisors, and self-raters) will assess performance and provide information regarding moderators. For example, attributions regarding cause of performance, self-esteem, amount of observational opportunity, as well as other possible factors would be measured.

Second, depending on the outcome of this research, further work on the ways to ameliorate rater bias would be useful. For example, if differential attributions are a source of conflict, the Performance Distribution Assessment (PDA), described by Bernardin and Beatty (1984), might provide a vehicle for increasing rater agreement. Specifically, the PDA attempts to take into account situational constraints. Research should also focus on identifying possible moderators of self–other agreement. For instance, impression management may be a variable that can account for self–supervisor and self–peer disagreement (Zerbe & Paulhus, 1987). Clearly, more research examining a variety of individual differences is in order.

In terms of practical implications, the results suggest that self-ratings will generally show only moderate correlations with ratings by others. This is particularly true for managerial/professional employees. Practitioners considering the use of self-ratings should be aware that there is liable to be much disagreement. Conversely, peer–supervisor ratings often (but not always) reflect relatively high rater agreement, even for managerial/professional jobs. For legal purposes, then, peer–supervisor ratings may be useful. Using raters from different levels may also help to develop consensus, eliminate biases, and perhaps in turn lead to greater acceptance by rates. Nevertheless, there is a plethora of other problems associated with using peer ratings (Latham & Wexley, 1982).
In conclusion, the present results show that peer–supervisor ratings demonstrate a relatively high correlation, whereas self–supervisor and self–peer ratings exhibit moderate correlations. Compared with the correlation between supervisor ratings and results-oriented outcomes (production, sales, etc.) (Heneman, 1986), the average correlation obtained here between supervisor and peer-ratings is substantially higher. Thus, from both an absolute and relative point of view, there is high convergence between observer ratings. The moderator analysis showed that correlations between self–peer and self–supervisor ratings are particularly low for managerial/professional workers, but not for peer–supervisor ratings. While these findings are consistent with an egocentric-bias explanation for rater disagreement, future research should test the alternative explanations described earlier in a more direct fashion than was possible here.

REFERENCES


(Eds.), Attribution: Perceiving the causes of behavior (pp.79–94). Morristown, NJ: General Learning Press.


STUDIES INCLUDED IN THE META-ANALYSIS


HARRIS AND SCHAUBROECK


