The Moderating Influence of Job Performance Dimensions on Convergence of Supervisory and Peer Ratings of Job Performance: Unconfounding Construct-Level Convergence and Rating Difficulty

Chockalingam Viswesvaryan
Florida International University

Deniz S. Ones
University of Minnesota, Twin Cities Campus

Frank L. Schmidt
University of Iowa

Meta-analysis was used to cumulate the correlations between supervisor and peer ratings for different dimensions of job performance. The general pattern in the data suggested that raters from the same organizational level disagree as much as raters from different levels. Methods were used to separate the effects of rating difficulty and lack of construct-level convergence on the correlation between supervisor and peer ratings. The authors found complete construct-level convergence for ratings of overall job performance, productivity, effort, job knowledge, quality, and leadership but not for ratings of administrative competence, interpersonal competence, and compliance or acceptance of authority. Higher rating difficulty was more strongly associated with lower mean observed peer-supervisor correlations than were construct-level disagreements between peers and supervisors. Implications for research and practice are discussed.

Performance ratings have traditionally played a central role in the measurement of job performance in industrial-organizational psychology. In their review, Lawshe and Balma (1969) found that 68% of validity studies used ratings. Lent, Aurbach, and Levin (1971) found that their review of 1,500 validation studies that 63% used ratings as the criterion. There is evidence that the use of ratings has not decreased over the years; if there has been any discernible trend, it has been an increasing use of ratings (Cascio, 1991; Landy, 1989; Marchese & Michlinsky, 1983). For example, Bernardin and Beatty (1984) found that 92% of the respondents to a survey indicated that they used ratings as their primary source of performance appraisal. Cleveland, Murphy, and Williams (1989) also found that the use of ratings is prevalent in organizations.

Because ratings of performance are widely used, they have been widely researched. Many researchers (e.g., Saad, Downey, & Lahey, 1980) have examined the psychometric properties of ratings such as distribution characteristics (e.g., central tendency and leniency—the skewness and kurtosis of the distribution), reliability (King, Hunter, & Schmidt, 1980; Rodolstein, 1990; Salgado & Moscoso, 1996; Viswesvaryan, Ones, & Schmidt, 1990), and validity (e.g., Church & Bracken, 1997). The effects of the rating format have been extensively researched, and consensus has emerged that the contribution of format differences to rating quality is limited (Landy & Farr, 1980). The effects of training (e.g., Borman, 1978; London & Smither, 1995) and training influences on cognitive processes have also been investigated (e.g., Hauenstein & Foti, 1989). Researchers have also examined the purposes for which ratings are used in an organization (Cleveland et al., 1989), the cognitive processes involved in ratings (e.g., DeNisi, Caftety, & Mلقino, 1984), and the influence of political climate in the organization on ratings (e.g., Kravitz & Balzer, 1992).

Ratings are subjective evaluations obtained from supervisors, peers, subordinates, self, or customers. The various uses of ratings can be broadly classified as administrative, feedback, or research. Of the five sources of evaluation, Cascio (1991, p. 81) stated that only supervisory and peer ratings are used for all three purposes to which ratings can be used. Most ratings are confined to supervisory ratings. For example, Lent et al. (1971) found that 93% of all ratings were supervisor ratings. (The remaining 7% of their database was based on peer ratings.) Bernardin and Beatty (1984) estimated that over 90% of the ratings used in the literature are supervisor evaluations. Furthermore, the traditional hierarchical structure of most organizations also increases the emphasis on supervisor ratings. In contrast, in recent years there has been a movement away from traditional hierarchical organizations toward matrix structures and teams or project-based organization of work (Norman & Zawacki, 1991), a trend that has increased the use of peer ratings and even of subordinate ratings.

This article focuses on the relationship between peer and supervisor ratings of job performance. It examines whether the correlations between ratings made by supervisors and those made by
peers is different for the different job performance dimensions rated (i.e., is the rating content a moderator of the convergence?). In examining the convergence in ratings between peers and supervisors, we distinguish between the effects of rating difficulty and construct-level convergence. Agreement between raters can be reduced by the absence of agreement on the nature of the construct to be rated, by the difficulty of rating a particular agreed on dimension, or by both. The correlation between peer and supervisor ratings may be reduced because peers and supervisors are rating different constructs or perceived dimensions of job performance (i.e., lack of construct-level convergence) because of differences in their understanding of the exact nature of the dimensions. That is, they may be actually rating somewhat different performance dimensions. Conversely, even when peers and supervisors are rating the same performance dimension (or construct), the correlation between peer and supervisor ratings of a performance dimension may be lower for one dimension than another because it is difficult to rate reliably, leading to lower supervisor-peer correlations. That is, if (for various reasons discussed later) some job performance dimensions are more difficult to rate, then the ratings on these dimensions will show lower interrater reliability, which will reduce peer–supervisor correlations. In this article, we refer to this effect or process as rating difficulty for the sake of brevity. The two effects, lack of construct-level convergence and rating difficulty, are confounded in the observed correlation between peer and supervisor ratings. By separating these two effects, it is possible for us to examine their relative importance. It is also possible for us to examine whether these two effects vary in strength across ratings of different dimensions of performance. For this article we use meta-analysis to examine these questions.

Although some recent literature seems to suggest that rater agreement does not depend on the organizational level of the raters (Mount, Judge, Scullen, Stysma, & Hezlett, 1998; Murphy, Cleveland, & Moher, 2001), most of the literature on 360° feedback systems clearly emphasizes the need for and importance of obtaining ratings from raters at different levels of the organization (e.g., peers and supervisors). Furthermore, prior meta-analysis of this literature focuses on peer–supervisor convergence on ratings of overall job performance. More fine-grained analyses that examine different performance dimensions and unconfound rating difficulty and construct-level convergence have the potential to inform both science and practice.

Rating Content as a Moderator of Peer–Supervisor Convergence

Several researchers have empirically examined the agreement between peer and supervisor ratings (e.g., Holzbach, 1978). Meta-analytic cumulations of this literature were undertaken by Harris and Schauhroeck (1988) and by Mabe and West (1982). Both of these meta-analyses examined only the average correlation between peer and supervisor ratings across the dimensions of job performance rated. However, most individual studies examining the convergence between peer and supervisor ratings have concluded that the level of agreement may depend on the dimension of job performance being rated. Many primary studies report data indicating that the correlation between peer and supervisor ratings depends on the content of ratings. For example, Albrecht, Glaser, and Marks (1964) found that the convergence between peer and supervisor ratings in rating sales ability was .74, whereas for rating interpersonal competence, the correlation was only .50. Borman (1974) found that the correlation between peer and supervisor ratings of responsibility was .05 whereas the correlation was .46 for ratings of job knowledge. Similarly, Holzbach (1978) found that the correlation between peer and supervisor ratings of quality was .20, but the correlation was .38 for ratings of productivity. These findings are representative of individual primary studies that empirically examined the convergence between peer and supervisor ratings for two or more dimensions of job performance.

A second line of evidence suggesting that the hypothesis that performance content domain moderates the convergence between peer and supervisor ratings comes from theoretical models that explicate the cognitive processes postulated to affect convergence between peer and supervisor ratings. For example, Wohlers and London (1989) presented some evidence that the rating difficulty of job performance dimensions affects the agreement between raters from different organizational levels. Specifically, they found evidence that more abstract performance dimensions, such as administrative competence, problem solving, leadership, and communication skills, produce less agreement than more concrete dimensions of job performance such as output level and accidents. Borman (1979, p. 419) found that raters rated rates significantly more accurately on some dimensions than on others and that for the most part, these differences were consistent across formats p. 419. Borman noted that this consistent dimension effect, even across a variety of formats (and across jobs and samples), may be due to factors inherent in the nature of the dimensions that makes them either difficult or easy for raters to rate. Borman (1979, p. 420) also suggested that "accuracy is highest on those dimensions for which actors provided the least ambiguous, most consistent performances, perhaps because they, as well as the student raters, understood those particular dimensions better than some of the other dimensions."

The hypothesis that certain dimensions of job performance are easier to evaluate than others is also found in the personality literature (e.g., Christensen, 1974) and the social psychology literature. For example, Bandura (1977) and Salancik and Pfeffer (1978) posited from a social information-processing framework that when there are no clear interpretable signs of behavior or where the standards of evaluation are ambiguous, peer–supervisor agreement will be lower, compared with when there are clear interpretable signs and the standards are unambiguous. This is also hypothesized to be true when certain dimensions of job performance have rare occurrence (low base rate) or have greater salience in memory (e.g., accidents). The literature, however, is ambiguous in identifying the specific dimensions that are difficult to rate. In fact, Wohlers and London (1989) presented data that suggested that raters considered all dimensions as less than average in difficulty of ratings. Rating difficulty is perhaps the result of several related mechanisms (opportunity to observe, existence of standards to evaluate performance, etc.). For this article, we used the cumulative empirical evidence to date to test a general hypothesis that some performance dimensions are more difficult to rate than others.

We also examined more specific hypotheses. Comparison of peer–supervisor ratings' convergence across different job performance dimensions provides a basis for an empirical test of the hypothesis (Borman, 1979; Wohlers & London, 1989) of a gradi-
Construct-Level Convergence Versus Rating Difficulty

A key question related to differences between peer and supervisor ratings is whether, for any given dimension of job performance, peers and supervisors are rating the same construct or performance dimension. If indeed they are, then the true-score correlation between peer and supervisor ratings is expected to be 1.00 (to within sampling error). That is, once the attenuating effects of measurement error are eliminated, agreement should be perfect. In a meta-analytic cumulation of the job performance ratings literature, Viswesvaran et al. (1996) reported mean estimates of supervisory interrater reliability and peer interrater reliability for different job performance dimensions. These reliability coefficients can be used to correct the mean observed peer-supervisor correlation for measurement error to determine whether corrected values are within sampling error of 1.00. In this connection, it is useful to examine the confidence interval around corrected correlations. Confidence intervals are computed as follows. The reliability values reported in Viswesvaran et al. (1996) are used to correct the endpoints of the confidence intervals for the mean observed correlations between peer and supervisor ratings for the effects of measurement error in each type of rating; the resulting values are the end points of the corrected correlation. If the underlying construct being rated is the same for both types of raters, then the upper limit of the confidence interval around the mean construct-level correlation should extend up to 1.00.

It is important to note here that the construct underlying peer ratings of a given dimension is defined as what is common across peers, and the construct underlying supervisor ratings of that same dimension is defined as what is common across supervisors. What is truly unique or idiosyncratic to a particular supervisor or peer (not shared with other supervisors or peers, respectively) is considered measurement error. This exactly parallels the way in which specific variance in items is assigned to measurement error in the computation of coefficient alpha in classical measurement theory.

Note that in this framework, when two peers (or two supervisors) view different performance episodes of the ratee to the extent that the different performance episodes viewed randomly sample from the same domain of behaviors (i.e., the same construct), multiple raters (e.g., two peers) can be conceptualized to be parallel forms. On parallel forms of tests, individual items are unique to each form; however, they randomly sample from the same construct domain. In the performance ratings domain, raters (e.g., two peers) viewing different specific episodes of behavior may be conceptualized similarly. True score on a performance dimension is defined in the measurement model as the mean rating over a large number of comparably qualified raters (i.e., randomly sampling or viewing behaviors tapping the same construct).

Given this measurement model, agreement between peers and supervisors can be reduced by two different processes: (a) measurement error (defined as the sum of random response error and rater-specific measurement error) and (b) construct-level divergence (defined as a lack of complete agreement between peers and supervisors on the nature of the construct to be rated). The relative level of rating difficulty across dimensions is reflected in amount or level of measurement error. Rating difficulty for supervisors is indexed by the intersupervisor reliabilities. Rating difficulty for peers is reflected in interpeer reliabilities.

For any given rated performance dimension, if the confidence interval reaches up to 1.00, then we can conclude that any lack of convergence between peers and supervisors may not be due to failure of the two types of raters to assess the same performance construct. In other words, we conclude that lack of perfect convergence in mean observed correlation is due to measurement error in both sets of ratings. In such cases, if the mean observed correlations differ across performance dimensions, then we conclude that this difference is likely due to the fact that there is more measurement error in some dimensions than others. The dimensions that are more difficult to rate are those that are less reliably rated (lower interrater reliability). This, in turn, results in smaller observed correlations between peer and supervisor ratings.

If a confidence interval does not include 1.00, then we conclude that one reason for lack of or low convergence between peer and supervisor ratings is that peers and supervisors may be rating somewhat different performance dimensions. That is, we conclude that convergence is lower than optimal for two reasons: (a) the biasing effects of measurement error in each set of ratings and (b) the underlying perceived performance dimension being assessed is somewhat different for peer and supervisor ratings. One reason that underlying constructs or performance dimensions might be somewhat different (although correlated) is potential differences in perspectives between peers and supervisors, as noted earlier. For example, if the estimated true-score correlation for a given dimension is only .50, then this would indicate that the effect of rating difficulty—which manifests itself in lower reliability—is merely to lower this correlation downward from its ceiling value of .50. The difference between .50 and the mean observed correlation is then the index of rating difficulty for this performance dimension. For example, if the mean observed correlation is .30, the estimated effect of measurement error is .50 − .30 = .20. For another performance dimension, this value could be higher or lower, indicating that there are variations across dimensions in rating difficulty.
Hence, an examination of the true-score correlations between peer and supervisor ratings is a diagnostic tool that can reveal information about the reasons for less than perfect convergence between peer and supervisor ratings. If convergence is low because of measurement error, then training programs aimed at reducing measurement error in rating for the affected sources (e.g., peers) might increase agreement; if construct-level convergence is low (i.e., unique source perspectives exist), then the implication may be that performance appraisal systems should be redesigned to include different perspectives by emphasizing 360° feedback systems. Recently, researchers have made attempts along these lines (Cheung, 1999; Moos et al., 1998) to disentangle construct-level convergence between peer and supervisor ratings from measurement error or rating difficulty. Our objective in this article is to use meta-analysis to confound the effects of rating difficulty and lack of construct-level convergence.

Method

Database

We searched the literature for articles that reported correlations between peer and supervisor ratings. To be included in the database, the ratings had to be based on actual job performance: interview ratings, assessment center ratings, and performance ratings in simulated exercises were excluded. We searched all articles (starting from the first issue of a journal) until the 1999 issue of the following 15 journals: Journal of Applied Psychology, Personnel Psychology, Academy of Management Journal, Human Relations, Journal of Business and Psychology, Journal of Management, Organizational Behavior and Human Decision Processes, Accident Analysis and Prevention, International Journal of Intercultural Relations, Journal of Vocational Behavior, Journal of Applied Behavioral Analysis, Human Resources Management Research Journal, Journal of Organizational and Occupational Psychology, Psychological Reports, and Journal of Organizational Behavior. A total of 40 studies were included in the database. These meta-analyses are indicated in the References with an asterisk.

Analysis

Viswesvaran et al. (1996) derived 10 dimensions of job performance by first extracting 486 job performance measures used in published articles over the years. The research reported here used Viswesvaran et al.'s dimensions of job performance to classify the performance dimensions found in the research literature. These 10 dimensions are characterized as follows: (a) overall job performance captures overall effectiveness, overall work reputation, or the sum of all individual dimensions rated; (b) productivity includes ratings of quantity of output or ratings of volume of work an individual produces; (c) effort focuses on the effort or motivation an individual expends in striving to do the job; (d) interpersonal competence reflects how well an individual gets along with others; (e) administrative competence reflects the proficiency an individual exhibits in coordinating his or her different roles in an organization; (f) quality assesses quality independent of quantity; (g) job knowledge assesses job-related knowledge and expertise an individual demonstrates; (h) communication competence reflects how well an individual communicates regardless of the content; (i) leadership measures an individual's ability to successfully bring out extra performance from others; and (j) compliance or acceptance of authority reflects favorableness of an individual's attitude toward organizational rules and regulations. More detailed explanations of these dimensions and illustrative examples are provided in Viswesvaran et al.'s article. Although arguments can be made that some of these performance dimensions are causally related (e.g., job knowledge and effort result in productivity), it should be noted that most models of job performance (e.g., Campbell's, 1990, model) treat them as performance dimensions. That is, models of the dimensionality of job performance—opposed to causal theories of job performance—do not explicitly distinguish between antecedent dimensions and consequence dimensions of job performance. This is also the case in the performance appraisal literature as a whole.

When a study contained conceptual replications (i.e., two measures of the same dimension were used in the same sample), linear composites with unit weights for the components were formed. Linear composites provide more construct valid. (Nunnally, 1978) and precise estimates than the use of the average correlation or the use of all correlations as separate data points (Hunter & Schmidt, 1990, pp. 457–463; Viswesvaran, Schmidt, & Ones, 1994).

The observed correlations between peer and supervisory ratings were first averaged across samples for each of the 10 performance dimensions. Because we had only two studies reporting a correlation between peer and supervisor ratings for the dimension of communication competence, we dropped that dimension. Thus, nine mean observed correlations were computed. Individual correlations were weighted by their respective sample sizes in computing the mean observed correlation. The sample size weighted standard deviation (SD) across the observed correlations was also computed and was used to compute the SE = SD/√k; where SD is the standard deviation of observed correlations and k is the number of studies. SE was used to compute confidence intervals for the mean r. If the confidence intervals for different job performance dimensions overlap, we cannot reject the hypothesis that rating difficulty is the same across the dimensions.

The mean observed correlation for each of the nine dimensions was next corrected using the interrater reliabilities for peer and supervisor ratings reported in Viswesvaran et al.'s (1996) article. These interrater reliabilities are shown in Table 1 (discussed later). Viswesvaran et al. had summarized the interrater reliabilities for the dimensions where four or more estimates were available. For supervisor ratings of administrative competence, interrater reliability in Viswesvaran et al.'s analyses was 42, but this estimate was based on less than four estimates and thus was not reported in that published article.

The mean observed correlation for each of the nine dimensions was corrected for interrater reliability of peer and supervisor ratings and the interrater reliability of supervisory ratings. Thus, the second step in our analysis was to estimate the true-score correlations between peer and supervisor ratings for each of the nine dimensions of job performance. In estimating the true-score correlation, when the two measures being correlated are provided by different raters, interrater-reliability coefficients are the appropriate reliabilities; interrater-reliability estimates are not appropriate (Schmidt & Hunter, 1996, 1999; Schmidt, Viswesvaran, & Ones, 2000; Viswesvaran et al., 1996; see Murphy & DeShon, 2000, for a contrary view). It is also important to note that interrater reliability corrects for both raters specific variance and random measurement error. Thus, no need to correct for intrarater reliability.

If the confidence interval around the mean corrected correlation extends to 1.00, then we cannot reject the hypothesis that supervisors and peers are rating the same job performance construct. Confidence intervals for mean construct-level correlations are produced by correcting the endpoints of the confidence intervals for mean observed correlations. Confidence intervals can be corrected for measurement error just as individual correlations or mean correlations can be corrected (Hunter & Schmidt, 1996, p. 120). In our tables (based on convention in psychological research), we report 95% confidence intervals, although interested readers may compute other confidence intervals of interest given reported values.

Results

Table 2 reports the results. All mean observed correlations in Table 2 are less than .50, indicating substantial disagreement.
between peers and supervisors. This disagreement may be because of rating difficulty (reflected in interpeer and intersupervisor reliability), dimension-level disagreements between peers and supervisors, or both.

To unconfound rating difficulty and construct-level convergence, we first examine the confidence interval around the estimated true-score correlation. If the interval includes 1.00, we cannot reject the hypothesis that peers and supervisors are rating the same performance construct. In such a case, we then evaluate rating difficulty by looking at the mean observed correlations for that performance dimension. If the confidence intervals around the estimated true-score correlations do not include 1.00, this suggests that peers and supervisors may be rating somewhat different constructs and rating difficulty may not be the whole explanation for less than perfect convergence.

For the nine performance dimensions summarized in Table 2, the confidence intervals around the true-score correlation do not include 1.00 for three performance dimensions—administrative competence, interpersonal competence, and compliance or acceptance of authority. Peers and supervisors may have somewhat different conceptualizations of these three dimensions. However, the underlying constructs are fairly highly correlated for peers and supervisors—.86 for interpersonal competence, .69 for administrative competence, and .78 for compliance or acceptance of authority. It is pertinent to note that the reliability reported for peer ratings of compliance or acceptance of authority (.71) is unusually high compared with the reliability estimates reported for other dimensions. For example, the range of interrater reliability for supervisors was .45–.63, and it was .33–.42 for peers (excluding the value of .71 for compliance). If we had used a value of .43 (the mean of interpeer reliabilities) instead of .71 for compliance ratings, then we would have concluded that peers and supervisors are rating the same construct when providing ratings of compliance with authority. On the basis of our decision rule, supervisors and peers appear to be rating the same construct when providing ratings of productivity, quality, job knowledge, leadership, overall job performance, and effort. For these six dimensions, confidence intervals for the mean construct-level correlations include 1.00.

<table>
<thead>
<tr>
<th>Job performance dimension</th>
<th>Supervisor</th>
<th>Peer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>k</td>
</tr>
<tr>
<td>Overall job performance</td>
<td>14,650</td>
<td>40</td>
</tr>
<tr>
<td>Productivity</td>
<td>2,015</td>
<td>19</td>
</tr>
<tr>
<td>Effort</td>
<td>2,714</td>
<td>24</td>
</tr>
<tr>
<td>Interpersonal competence</td>
<td>3,006</td>
<td>31</td>
</tr>
<tr>
<td>Administrative competence</td>
<td>1,120</td>
<td>9</td>
</tr>
<tr>
<td>Quality</td>
<td>1,225</td>
<td>10</td>
</tr>
<tr>
<td>Job knowledge</td>
<td>14,072</td>
<td>20</td>
</tr>
<tr>
<td>Leadership</td>
<td>2,171</td>
<td>20</td>
</tr>
<tr>
<td>Compliance or acceptance of authority</td>
<td>905</td>
<td>8</td>
</tr>
</tbody>
</table>

Note. These values are summarized from Viswesvaran et al.'s (1996) article. Values are for ratings produced by a single rater. N = total sample size across all estimates that were averaged; k = number of samples averaged.

<table>
<thead>
<tr>
<th>Job performance dimension</th>
<th>k</th>
<th>N</th>
<th>r̂</th>
<th>95% CI (r̂)</th>
<th>̂p</th>
<th>95% CI (̂p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall job performance</td>
<td>31</td>
<td>6,253</td>
<td>.46</td>
<td>.41–.51</td>
<td>.98</td>
<td>.88–1.00</td>
</tr>
<tr>
<td>Productivity</td>
<td>10</td>
<td>2,112</td>
<td>.40</td>
<td>.35–.45</td>
<td>.91</td>
<td>.79–1.00</td>
</tr>
<tr>
<td>Effort</td>
<td>13</td>
<td>1,463</td>
<td>.48</td>
<td>.41–.56</td>
<td>.99</td>
<td>.84–1.00</td>
</tr>
<tr>
<td>Interpersonal competence</td>
<td>19</td>
<td>2,494</td>
<td>.38</td>
<td>.32–.43</td>
<td>.86</td>
<td>.74–.98</td>
</tr>
<tr>
<td>Administrative competence</td>
<td>6</td>
<td>1,164</td>
<td>.34</td>
<td>.24–.45</td>
<td>.69</td>
<td>.48–.90</td>
</tr>
<tr>
<td>Quality</td>
<td>3</td>
<td>374</td>
<td>.35</td>
<td>.12–.58</td>
<td>.68</td>
<td>.24–1.00</td>
</tr>
<tr>
<td>Job knowledge</td>
<td>9</td>
<td>1,685</td>
<td>.36</td>
<td>.27–.45</td>
<td>.86</td>
<td>.66–1.00</td>
</tr>
<tr>
<td>Leadership</td>
<td>7</td>
<td>544</td>
<td>.41</td>
<td>.30–.52</td>
<td>.91</td>
<td>.66–1.00</td>
</tr>
<tr>
<td>Compliance or acceptance of authority</td>
<td>6</td>
<td>768</td>
<td>.49</td>
<td>.38–.60</td>
<td>.78</td>
<td>.61–.95</td>
</tr>
</tbody>
</table>

Note. The performance dimensions were taken from, and described in, Viswesvaran et al.'s (1996) article. k = number of samples included in the meta-analysis; N = total sample size across the k samples; r̂ = sample size weighted mean of the observed correlations; 95% CI (r̂) = 95% confidence interval constructed around the mean observed correlation; ̂p = estimated true-score correlation; 95% CI (̂p) = 95% confidence interval for ̂p.
The mean observed peer-supervisor correlations are less than .50 for all nine dimensions. Combined with the finding that the confidence intervals for the mean correlation corrected for inter-peer and intersupervisor reliability (i.e., corrected for rating difficulty) included 1.00 for six of the nine dimensions, this pattern suggests that rating difficulty (as reflected in level of measurement error) is the major factor depressing peer-supervisor convergence. Is this observed correlation convergence moderated by the rating context? The confidence intervals around the mean observed correlations overlap for the six dimensions for which the confidence interval on the true-score correlation includes 1.00, suggesting that moderating influences, if any, are weak. Confidence intervals on observed correlations also overlap for the three dimensions for which the true-score correlations may be less than 1.00.

To some extent, the present findings allow us to examine the hypotheses about specific performance dimensions found in the literature and discussed earlier. Of the performance dimensions hypothesized in the literature to be relatively easy to rate, three are contained in Table 2: productivity, effort, and job knowledge. For all three of these dimensions, the confidence interval around the mean construct level includes 1.00, suggesting that supervisors and peers may be rating the same construct. The mean true-score correlation for these dimensions is .92. The average observed correlation between supervisor and peer raters is .41. Of the performance dimensions hypothesized in the literature to be relatively difficult to rate, two are contained in Table 2: administrative competence and leadership. (A third, communications skills, was dropped because of lack of data.) For one of these, leadership, the true-score correlation confidence interval reaches up to 1.00, and the mean true-score correlation is large (.91). Also, for this dimension the mean observed score correlation between supervisor and peer raters is .41, the same as the mean value for the low rating difficulty dimensions of productivity, effort, and job knowledge. Hence, there is no evidence that leadership is more difficult to rate than these low rating difficulty performance dimensions. The other performance dimension hypothesized in the literature to be relatively difficult to rate is administrative competence. For this dimension, the mean construct-level correlation is .69, and the confidence interval does not reach up to 1.00, suggesting that supervisor and peer raters are not rating the same construct. In addition, the mean observed correlation between supervisor and peer raters is .34, the lowest value in Table 2. However, for both observed and construct-level correlations, the confidence intervals overlap those of other dimensions and, in particular, overlap the confidence intervals of the three dimensions hypothesized to be relatively easy to rate. Hence, these findings do not provide strong support for the hypothesis that administrative competence is more difficult to rate than other dimensions, including those hypothesized to be easy to rate. For the remaining performance dimensions—overall job performance, interpersonal competence, quality, and compliance or acceptance of authority—the literature contained no specific hypotheses about rating difficulty.

The mean observed correlations for peer-supervisor convergence reported here can be compared with the interrater reliabilities reported in Viswesvaran et al.’s (1996) article for these nine dimensions for supervisors and peers. These interrater reliabilities are for single raters and are presented in Table 1. Interrater reliabilities are higher for supervisors (M = .55) than for peers (M = .43), indicating there is more measurement error in peer ratings than in supervisor ratings. Most of the mean observed convergence correlations for the nine dimensions in Table 2 (M = .41) are intermediate between supervisor and peer interrater reliabilities, again suggesting that measurement error (rating difficulty) is the primary determinant of level of peer-supervisor convergence.

This question can be examined in another way. The average construct-level correlation in Table 2 is .85, whereas the average observed score correlation is .41. This means that, on average across the performance dimensions, measurement error reduces the observed correlation between supervisor and peer raters from .85 to .41, a reduction of 44 correlation points. In contrast, lack of complete construct convergence reduces the mean supervisor-peer construct-level correlation from 1.00 to .85, a much smaller reduction. Hence, it is clear that measurement error is the major factor reducing correlations between supervisor and peer ratings. The effect of measurement error is, on average, almost three times larger than the impact of lack of complete construct convergence (.44/15 = 2.93).

If we omit the two dimensions with the lowest mean supervisor-peer construct-level correlations—administrative competence (.69) and quality (.68)—these figures are even more striking. The mean true-score correlation is .90, and the mean observed correlation is .41. Hence, measurement error reduces the mean supervisor-peer correlation even more: .90 − .41 = .49. However, the reduction in the mean construct-level correlation is only 1.00−.90 = .10.

For six of the nine performance dimensions, we cannot reject the hypothesis that the mean true-score correlation is 1.00. However, we can evaluate the data in Table 2 from a more holistic perspective. In this connection, it is striking that none of the estimates of mean true-score correlations are 1.00 or exceed 1.00. Suppose that the real (population) value of every true-score correlation were in fact 1.00. Then half the estimated mean true-score correlations should be (at least slightly) less than 1.00 and half should be (at least slightly) above 1.00—because of sampling error in the estimates (Schmidt & Hunter, 1999, cf. also Hunter & Schmidt, 1990, chap. 9). The expected average would be exactly 1.00. Instead, none of the values exceeds 1.00, and the average value is only .85 (.90 if we exclude the two performance dimensions with the lowest mean true-score correlations). This pattern of results suggests that supervisors and peers are, on the average, rating similar but not exactly identical performance constructs. The lack of perfect agreement between supervisor and peer raters is mostly due to measurement error. However, a small part of it is due to lack of complete construct convergence between supervisor and peer raters. This is the average pattern of findings. However, for some specific dimensions, the supervisor-peer construct convergence may be complete. For example, for overall job performance, the mean true-score correlation is .98, and for the effort dimension, it is .99.

The variation in mean interrater reliabilities across performance dimensions may be viewed as reflecting differences in rating difficulty for performance dimensions. An important question is whether these differences in rating difficulty generalize across rating sources. Are dimensions that are difficult for peers to rate also difficult for supervisors to rate? We correlated the mean interrater reliabilities of supervisors and peers across the nine performance dimensions and obtained a correlation of .21. Hence, the dimensions that are the most difficult to rate reliably are not the
same for supervisor and peer raters. However, this correlation is computed on a small sample, and, therefore, any inference on the basis of it has to be tentative.

Discussion

Ratings are widely used in performance appraisals, and peers and supervisors are the two most common sources of ratings. For this article, we integrated research examining the convergence (correlation) between peer and supervisor ratings. Our findings indicate that the moderating effect of the rating content on the convergence between peer and supervisor ratings is not as strong as it is implied by some cognitively based hypotheses (e.g., evaluation difficulty) proposed to explain the rating processes. Our findings are also consistent with the conclusions of Mount et al. (1998) that source effects (e.g., peer, supervisor) are not strong in ratings.

Our findings indicate that peer-supervisor convergence across dimensions apparently does not vary much by job performance dimension at either the level of observed correlations or the level of construct-level correlations. This calls into question the importance of processes postulated to moderate convergence (cf. Borman, 1974, 1979; Wohlers & London, 1989). Variability in peer-supervisor convergence across dimensions is a necessary but not sufficient condition to support the mechanisms postulated to moderate convergence. The meta-analytic cumulation reported here indicates that there is only limited variability in peer-supervisor convergence across dimensions. Thus, the condition (substantial variability in peer-supervisor convergence across dimensions) necessary for inferring support for the mechanisms postulated to moderate convergence appears to be only weakly satisfied.

Another interesting point that emerged from our results that could profit from future research scrutiny is the high reliability of compliance ratings compared with ratings of other dimensions. Perhaps behaviors related to compliance and accepting authority may be regarded as extensions of mores and norms encountered in society at large, providing raters with a common frame of reference. Yet another issue is the question of the construct validity of ratings. If interrater reliabilities are as low as we found in the cumulative literature, then to what extent can organizations, and we as scientist-practitioners, rely on a single supervisor’s ratings to validate our interventions? It is true that when an organization bases its decisions on the ratings of a single supervisor, there is a lot of unreliability in that measure.

Having said that, we also point out that reliability and construct validity, though related, are not the same. The fact that the reliability is only .50 implies that the construct validity of the observed score (i.e., the correlation between the observed scores of a measure and the underlying construct) has to be .71 (>.50) or less. (Reliability is the ratio of true to observed variance, and the square root of the reliability coefficient is the correlation between observed scores and the scale’s underlying true scores [Nunnally & Bernstein, 1964]). The correlation between the observed scores and the construct scale is intended to measure cannot be higher than the correlation between observed scores and true scores of the scale (the square root of the reliability). Therefore, if the reliability of ratings is .50, then the implication is that the construct validity cannot be higher than .71. This is still a high and respectable index of construct validity. The construct validity of ratings has been researched for many years, and even a brief summary is beyond the scope of this article.

One can argue further that because the concept of interrater reliability rests on the premise that one is computing the correlation between parallel raters, interrater reliability for supervisor ratings is problematic because there is only one true supervisor. To compute the interrater reliability of supervisor ratings, a second person (in most cases, the supervisor’s supervisor) is located to provide second ratings. Is the second rater a parallel measure of the first, true, supervisor? If not, is interrater reliability underestimated for supervisors? (Note that in most organizations, individuals would have reported to more than one supervisor over time. Employees rotate among supervisors, and, as such, estimating interrater reliability may not be a serious problem.)

There are two points to note in addressing this concern. First, as seen in Table 1, the average interrater reliability for supervisors is higher than that reported for peers (where presumably two parallel raters are available). It is more likely that an individual has two peers but only one immediate supervisor. Thus, intersupervisor reliability is computed by having a supervisor’s supervisor, or a stand-in or a rotating supervisor, provide ratings in addition to the immediate supervisor. However, the computation is relatively more straightforward for interpeer reliability because individuals are likely to have two peers. If the various factors are influencing interrater reliabilities, then they are more likely to affect supervisors than peers. Thus, supervisor reliability should be lower than estimated peer interrater reliability. However, the cumulative evidence shows that interrater reliability is lower for peers than it is for supervisors. This seems to be evidence against the hypothesis that interrater reliability for supervisory ratings is underestimated in the literature because there are no two parallel supervisors (i.e., the supervisor is unique to each ratee).

Second, there is another factor—leniency—to take into account. Raters differ in leniency. Typically, in computing the interrater reliability, some individuals will be rated by two raters, others by another set of two raters, and so forth. Thus, rater main effects (leniency) are confounded in Rater X Ratee interactions. An argument can be made that the interrater estimate is therefore affected. However, the reliability estimate used should match the type of data being correlated in the real world. In almost all instances, we have different supervisors rating different employees. These ratings are then correlated with test or predictor data. Thus, the reliability coefficient used to correct the correlations should also include the rater main effect as error. The interrater reliability estimates that we have here accurately reflect this real-world dynamic of obtaining data—and hence are the appropriate ones to use in making the corrections.

At this point, we state our position on another issue that was raised first by the editor-in-chief of this journal, the initial action editor of this article, and subsequently by the associate editor who served as the action editor. Both claimed that our framework is valid only if one accepts our definitions of true score and error variance. Specifically, both argued that rater-specific variance need not be construed as error. They suggested that raters may observe different episodes of behavior, resulting in true disagreements between raters that should not be interpreted as measurement error (see also Murphy & Cleveland, 1995). Even though two raters may observe different incidents of performance, as long as the different incidents of performance viewed tap into the same
performance dimension (sample the same domain), raters can be viewed as parallel raters. Under the classical-measurement model, rater-specific variance is measurement error just as item-specific variance is measurement error in the computation of coefficients of equivalence and alpha. A more detailed discussion of this issue is available elsewhere (Murphy & DeShon, 2000; Schmidt et al., 2000). Here, we state that our framework is built only on the principles of classical-measurement theory, and, to the extent these principles are inappropriate, our framework and conclusions will also be inappropriate. However, it is our position that classical-measurement theory is appropriate for use with ratings of job performance (Schmidt et al., 2000).

In this study, we focus on whether the confidence intervals for any two performance dimensions overlap. A reviewer suggested that instead we should compute the confidence interval around the difference between the correlations for each pair of performance dimensions. Although this could be done, it would result in 36 different confidence intervals (vs. the nine in our analysis) and would not affect the conclusions. However, the information provided in Table 2 is sufficient to allow readers desiring such confidence intervals to compute them.

This research makes several substantive and methodological contributions. First, this article presents the first meta-analytic cumulation of peer–supervisor convergence for different dimensions of job performance. Previous meta-analytic cumulations (e.g., Harris & Schaubroeck, 1988) have not presented results separately for different dimensions. Comparing the convergence for different dimensions is a necessary first step in investigating underlying processes in ratings, such as rating difficulty (cf. Borman, 1979), that have been postulated in the personality, performance appraisal, and social information-processing literatures. A second unique feature of this study is the exclusive use of interrater reliability coefficients for correcting correlations between peer and supervisor ratings. Previous meta-analytic cumulations have included both intrarater and interrater reliability estimates in the reliability distributions, creating a downward bias in the estimated true-score correlations. Appropriate reliability distributions are required to accurately estimate true-score correlations (Schmidt & Hunter, 1996; Viswesvaran et al., 1996). Finally, the effects of construct-level convergence and rating difficulty were unconfounded in this study.

A limitation in this meta-analytic cumulation is the small number of studies in some categories. Future research is needed to address this concern (and thereby reduce the second-order sampling error potentially present in these analyses). In contrast, the smallest analysis in Table 2 is based on N = 374, a sample size that may consider large in published individual empirical studies. It would also be worthwhile to examine other potential moderators of peer–supervisor convergence, such as whether the ratings were obtained for administrative or research purposes, length of exposure, and so forth (see Wherry & Bartlett, 1982, for a discussion of the different sources of variance influencing ratings). However, previous meta-analytic investigations of reliability estimates in many different domains (Churchill & Peter, 1984; Peterson, 1994) have not found strong evidence of such moderator variables.

In this article, we have focused on peer and supervisor ratings of different dimensions of job performance. Future research should focus on other sources of ratings: customers, subordinates, and self. More primary studies are required that examine the correlation (i.e., the convergent validity) for different dimensions of job performance for customer, subordinate, and self-ratings. For example, Mount et al. (1998) reported mean observed correlations between subordinate ratings and peer ratings and between subordinate and supervisor ratings. We see the efforts of Mount et al. as a valuable step in reaching generalizable conclusions about the convergence of subordinate ratings with supervisor and peer ratings.

The findings reported in this article have implications for performance appraisal systems. Industrial–organizational psychologists invested much effort in the 1960s and 1970s in trying to identify the best rating scale formats. As noted earlier, this research eventually concluded that rating scale format had minimal impact on rating quality (Landy & Farr, 1980). A current belief is that the factors affecting ratings (e.g., rating difficulty) differ across dimensions. Our findings support the conclusion that agreement between ratings made by peers and supervisors differs little across rated dimensions of job performance.

Supporting the belief that the factors affecting ratings (e.g., rating difficulty) vary across dimensions was the belief that the cognitive processes that underlie ratings and the political processes affecting performance appraisal are dependent on the dimension rated. Our findings indicate that we need to focus on processes, such as halo, that explain the general rating process. These processes do not appear to depend very much on the particular content of performance dimensions. The empirical findings we report here suggest that the rating processes and influences on the performance appraisal systems may not be specific to the job performance dimension being rated. If so, it may be more profitable for future research on performance appraisal to focus on the general process of performance appraisal.

References

References marked with an asterisk indicate studies included in the meta-analysis.


Received November 8, 1999
Revision received July 17, 2001
Accepted July 17, 2001