Industrial and Organizational Psychology, **1** (2008), 161–166.

COMMENTARIES

On the Relationship Between Job Performance and Ratings of Job Performance: What Do We Really Know?

DAVID J. WOEHR
The University of Tennessee

Murphy (2008) provides a clear and concise statement of various models of job performance ratings. Indeed, there is much to be gained, both conceptually and practically, from such a systematic analysis and articulation. Further, Murphy's ultimate conclusion, that an increased focus on better "climates" for performance rating will result in better performance ratings, certainly represents a positive next step for the improvement of performance evaluation. The organizational context in which ratings occur is definitely a vital consideration with respect to the quality of the performance ratings obtained. Future research should certainly focus on defining those aspects of the organizational context that are most important for high-quality performance ratings as well as methodologies for evaluating the climate for performance appraisal in organizations.

Yet, Murphy builds this analysis on a troubling cornerstone. Specifically, as emphasized in the title and throughout the paper, the starting point is the "weak relationship between job performance and ratings of job performance." This premise is particularly troubling in several of its implications. Three potential implications stand out: (a) there is solid evidence that the relationship between job performance and ratings of job performance is weak, (b) research on performance appraisal and rating interventions to date has had little practical impact on the relationship between ratings and performance, and (c) future work focusing on organizational development (OD) type interventions will lead to a demonstrated improvement in the relationship between ratings and performance. These implications warrant further examination.

> Implication 1: There is solid evidence that the relationship between performance and job performance is weak.

Arguably, this is one of the most firmly entrenched tenets of Industrial and Organizational (I–O) psychology—subjective ratings of job performance are poor measures of actual performance. But I–O psychology also prides itself on being an empirical, data-based science. So statements throughout Murphy's analysis such as "… most reviews of performance appraisal research … *suggest* that the relationship between job performance and ratings of job performance is *likely to be weak, or at best uncertain*" (p. 151, italics added) or "… but it does *seem* clear that the relationship between job performance and ratings of job performance is *not likely to be strong*" (p. 151, italics added) are troubling. "Likely to be" or "seem to be"

Correspondence concerning this article should be addressed to David J. Woehr. E-mail: djw@utk.edu
Address: Department of Management, The University of Tennessee, 411 Stokely Management Center, Knoxville, TN 37796
David J. Woehr, Department of Management, The University of Tennessee.

https://doi.org/10.1111/j.1754-9434.2008.00031.x Published online by Cambridge University Press

are a far cry from "have been shown to be" or "evidence demonstrates." Given all the research to date, why can we not be more definitive on this point? At a conceptual level, the answer lies in the fact that we do not, and probably cannot, ever know the level of true job performance. Thus, all efforts to date rely on indirect sources of evidence with respect to the relationship between ratings and job performance. And in fact, traditional evidentiary bases for the supposition that this relationship is weak are seriously flawed sources of evidence and thus really say very little as to the nature of this relationship.

Criteria for evaluation of criterion measures has been a key concern in the I–O literature for most of the field's history (Austin & Crepin, 2006). The key difficulty is that it is practically impossible to separate the "what" from the "how." That is, it is not possible to directly assess the relationship between ratings and actual performance in that it is impossible to operationally define actual performance without using ratings or some other potentially equally problematic method. Rather, we are forced to make suppositions about the relationship between ratings and performance based on other "evidence." What is this other evidence? Four general categories are reflected in the literature. These are (a) psychometric rating "errors" (i.e., rating distributional outcomes), (b) "rating accuracy," (c) information processing errors, and (d) interrater agreement.

Psychometric Rating Errors

One of the earliest set of criteria for the evaluation of performance ratings was an examination of the psychometric and/or distributional properties of the ratings. The pervasiveness of negatively skewed, range-restricted, and moderately to highly intercorrelated performance ratings is well documented (Cooper, 1981; Landy & Farr, 1980; Saal, Downey, & Lahey, 1980) and has been interpreted to indicate the presence of leniency, central tendency, and halo "rating errors," respectively (Saal et al.). These psychometric errors have been interpreted as evidence that ratings contain substantial amounts of both systematic and nonsystematic performance irrelevant information. Yet, it is also widely recognized that these psychometric characteristics do not necessarily reflect "error" (Murphy & Balzer, 1989). That is, to the extent that rating distributions reflect actual performance distributions, these findings are not indicative of rating error but of true performance. So to what extent do psychometric properties of ratings tell us something about the relationship between actual performance and performance ratings? The answer is very little.

Rating Accuracy

As highlighted above, the use of psychometric rating errors presents a fundamental problem. That is, it is really impossible to know the extent to which the observed rating distributions reflect the actual performance distributions (Levy & Williams, 2004). Thus, a good bit of performance appraisal research has focused on rating accuracy as an alternative criterion.

Although there are a variety of operationalizations of rating accuracy, common to all is the requirement of a standard to which ratings are compared (Roach & Gupta, 1992; Sulsky & Balzer, 1988). In essence, the more similar the actual ratings are to the standard, the more accurate. However, rating accuracy is really a misnomer in that ratings are not compared with actual job performance but to another set of ratings (usually provided by a set of "expert" raters observing and rating performance under optimal conditions). Yet, what evidence do we have that the "expert ratings" actually reflect "true performance" or are "better" than some other set of ratings? None, and thus we come full circle and all we can definitely conclude is how one set of ratings compare with another. Another major limitation of rating accuracy criterion is that they require a controlled performance for which expert ratings can be developed and that same exact performance must subsequently be presented to raters. This limitation generally constrains the use of rating accuracy to controlled laboratory settings, which

https://doi.org/10.1111/j.1754-9434.2008.00031.x Published online by Cambridge University Press

raise important ecological validity questions (Dipboye, 1990; Ilgen, Barnes-Farrell, & McKellin, 1993).

Information Processing Errors

Given the problems associated with the use of psychometric rating errors and rating accuracy as a criterion for the evaluation of performance ratings, the literature moved to a consideration of performance rating as social information processing (DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981). Here, the general idea was that the performance rating process is best viewed as a specific type of the more general person perception or attribution process. As such, it is presumed that ratings are subject to the wide range of information processing errors demonstrated in the social judgment literature. Because of this, the social judgment literature paints a fairly dismal picture of human social reasoning. And by extension, performance ratings, as a type of social judgment, are assumed to be equally problematic. That is, social information processing errors demonstrated in the laboratory represent inaccurate judgments and thus, actual performance ratings are *likely* the product of flawed judgment processes. However, a serious limitation of the social judgment literature in general and the job performance judgment literature in particular is that it is almost completely laboratory based. As noted by Funder (1987), such laboratory research has very limited applicability to real-life decision making. Specifically, Funder argues:

> Although errors can be highly informative about the *process* of judgment in general, they are not necessarily relevant to the *content* or accuracy of particular judgments, because errors in the laboratory may not be mistakes with respect to a broader, more realistic frame of reference and the processes that produce such errors might lead to correct decisions and adaptive outcomes in real life. (p. 75)

In short, although the performance information processing literature may indicate that processing errors may occur, it does not provide a very good indication of the extent to which such errors do occur in applied settings nor, more importantly, the extent to which they reduce the relationship between ratings and actual job performance.

Interrater Agreement

Driven to a large extent by the popularity of multisource feedback systems, there has been an increasing focus on the level of agreement both within and between ratings sources. As noted by Murphy and Viswesvaran, Ones, and Schmidt (1996), the "Achilles Heel" of these systems is consistent findings of a lack of agreement across different rating sources. But what does this lack of agreement say about the relationship between job performance and ratings? The traditional psychometric argument is that consistency (across measurements or sources) is necessary but not sufficient condition for validity. Thus, a lack of consistency precludes validity. However, there is a long history of debate focusing on whether or not there should be agreement across sources (Borman, 1974; Murphy & DeShon, 2000; Schmidt, Viswesvaran, & Ones, 2000). That is, lack of agreement across sources may reflect true differences resulting from differences in perspectives or opportunities to observe performance. Yet, even this argument may be somewhat moot. Recent research suggests that findings of low levels of agreement across rating sources may be largely artifactual. Specifically, LeBretton, Burgess, Kaiser, Atchley, and James (2003) present a convincing case that estimates of interrater agreement based on intraclass and Pearson correlations are severely attenuated because of restriction of range in job performance and thus represent substantial underestimates of interrater agreement. LeBretton et al. demonstrate that non–correlation-based methods of assessing interrater agreement indicate relatively high levels of agreement. So what does this tell us about the validity of job performance ratings. Again, not much.

So in sum, what can we definitively say about the relationship between job performance and ratings of job performance? It

https://doi.org/10.1111/j.1754-9434.2008.00031.x Published online by Cambridge University Press

may be possible to make a logical, theoretical case that this relationship is *likely to be weak*. But we cannot conclude based on data that this is actually the case. Thus, the premise that the relationship between performance and ratings of performance is weak is supported by flawed evidence and is premature at best.

> Implication 2: Research on performance appraisal and rating interventions to date has had little practical impact on the relationship between ratings and performance.

Perhaps the most troubling implication raised by Murphy is that interventions aimed at improving the relationship between job performance and ratings of performance have largely been ineffective. Here, it is important to distinguish the interventions and strategies proposed and evaluated from the criteria used to evaluate them. If the criteria are problematic, then it becomes impossible to evaluate whether or not the interventions are effective. To date, performance appraisal interventions have largely focused on rating scale development and rater training. Both of these intervention strategies have been evaluated in terms of the set of criteria discussed above (i.e., psychometric properties, information processing errors, interrater agreement, and rating accuracy). Can we conclude from this research that rating scale development and rater training are unimportant or ineffective? I believe not. First, it is not at all clear that the criteria used provide an adequate evaluation of the impact of the interventions. Second, it is important to clearly understand the nature of the findings to date. For example, although much of the rating scale literature indicates that specific scale format may not lead to major differences in rating outcomes, it is all predicated on the use of job-relevant professionally developed scales. So although it may not matter if one uses behaviorally anchored ratings scales, behavioral observation scales, or Likert-type scales, it does matter that the scales used are based on a thorough job analysis and incorporate clear behaviorally based definitions of the constructs to be evaluated. Similarly, the rater training literature presents a consistent picture that providing raters a clear and consistent explanation of what and how they are supposed to rate along with practice and feedback doing so, greatly facilitates the rating process (Woehr & Huffcutt, 1994).

The main point here is that it is premature at best and erroneous at worst to conclude that, on the basis of the research to date examining the impact of rating scale and rater training interventions on the set of criteria described above, these interventions are ineffective and do not impact the relationship between ratings and job performance.

> Implication 3: Future work on improving the climate for performance appraisal will lead to a demonstrated improvement in the relationship between job performance and ratings of job performance.

Murphy ultimately concludes that OD interventions are more likely to improve the quality of performance appraisals in organizations than are traditional scale development and training interventions. There is no denying the potential importance of such OD-based interventions. Certainly, rating outcomes are determined as much, if not more so, by rater motivation to provide high-quality ratings as by the ability to do so. It is also clear that this type of intervention has largely been neglected in the performance appraisal literature. But will these interventions lead to a demonstrated increase in the relationship between job performance and ratings of job performance? Given the aforementioned issues with traditional rating criteria, the answer is probably no. That is, will such interventions change the distributional properties of ratings, increase rating accuracy, reduce potential information processing/attributional biases, or increase interrater agreement? Probably not.

In conclusion, I would like to emphasize three points. First, before we seek to explain the weak relationship between job performance and ratings of job performance, we need to know that this relationship is in fact weak. Murphy contends that the three

https://doi.org/10.1111/j.1754-9434.2008.00031.x Published online by Cambridge University Press

models presented "all agree on one essential point—i.e., that the relationship between job performance and ratings of job performance is likely to be weak" (p. 157). This is not strictly true. Rather, all the models presented suggest that performance ratings are potentially a function of multiple factors, only one of which is job performance. None of the models provide an indication of the relative magnitude of the impact of these factors. For this, we must turn to the existing evidence. But traditional evidentiary bases for the supposition that this relationship is weak are seriously flawed and thus really say very little as to the nature of this relationship. Perhaps it is time that we revisit the evidence underlying the premise that the relationship between job performance and ratings of performance is as dismal as generally purported.

Second, we must be cautious with respect to statements about the value or effectiveness of existing interventions. Given the potential limitations associated with traditional evidentiary bases for evaluating performance ratings, it is inappropriate to conclude that scale development and rater training are unimportant or ineffective. Further, claims that we know all we need to know about these interventions and relatedly calls for research moratoriums are similarly unwarranted.

Finally, we should also be cautious about setting unrealistic expectations with respect to the impact of OD-based interventions. Future research should certainly focus on defining those aspects of the organizational context that are most important for high quality performance ratings. Yet, if we continue to operationally define "high quality" as we always have, it is not likely that we will see any stronger evidence on the relationship between ratings and performance than we have to date. Along these lines, perhaps it is time to acknowledge that we really do not know, and likely cannot ever know, what the true relationship between job performance and ratings of job performance is. Rather, we should work to refine existing interventions as well as develop new interventions that are as conceptually grounded and methodologically rigorous as possible. Clearly, both rater ability and motivation are important determinants of rating outcomes. Scale development and rater training interventions have largely focused on ability while ignoring motivation. It is important that we examine interventions focused on rater motivation. However, in doing so, it is equally important that we do not forget or downplay the ability component.

References

Austin, J. T., & Crepin, T. R. (2006). Problems of criteria in industrial and organizational psychology: Progress, problems, and prospects. In W. Bennett, Jr., C. E. Lance, & D. J. Woehr (Eds.), *Performance measurement: Current perspectives and future challenges*. Mahwah, NJ: Lawrence Erlbaum Associates.

Borman, W. C. (1974). The rating of individuals in organizations: An alternative approach. *Organizational Behavior and Human Performance, 12*, 105–124.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*, 218–244.

DeNisi, A. S., Cafferty, T., & Meglino, B. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Decision Processes, 33*, 360–396.

Dipboye, R. L. (1990). Laboratory vs. field research in industrial and organizational psychology. *International Review of Industrial and Organizational Psychology, 5*, 1–34.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*, 127–148.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101*, 75–90.

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Process, 54*, 321–368.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72–107.

LeBretton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Orgnaizational Research Methods, 6*, 80–128.

Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management, 20*, 881–905.

Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 148–160.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*, 619–624.

https://doi.org/10.1111/j.1754-9434.2008.00031.x Published online by Cambridge University Press

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 73–900.
Roach, D. W., & Gupta, N. (1992). A realistic simulation for assessing the relationships among components of accuracy. *Journal of Applied Psychology*, *77*, 196–200.
Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428.
Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*, 901–912.
Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*, 497–506.
Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574.
Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Organizational and Occupational Psychology*, *67*, 189–205.

https://doi.org/10.1111/j.1754-9434.2008.00031.x Published online by Cambridge University Press