Industrial and Organizational Psychology, **1** (2008), 148–160.

FOCAL ARTICLE

Explaining the Weak Relationship Between Job Performance and Ratings of Job Performance

KEVIN R. MURPHY
The Pennsylvania State University

Abstract

Ratings of job performance are widely viewed as poor measures of job performance. Three models of the performance–performance rating relationship offer very different explanations and solutions for this seemingly weak relationship. One-factor models suggest that measurement error is the main difference between performance and performance ratings and they offer a simple solution—that is, the correction for attenuation. Multifactor models suggest that the effects of job performance on performance ratings are often masked by a range of systematic nonperformance factors that also influence these ratings. These models suggest isolating and dampening the effects of these nonperformance factors. Mediated models suggest that intentional distortions are a key reason that ratings often fail to reflect ratee performance. These models suggest that raters must be given both the tools and the incentive to perform well as measurement instruments and that systematic efforts to remove the negative consequences of giving honest performance ratings are needed if we hope to use performance ratings as serious measures of job performance.

The measurement of job performance has long been recognized as one of the significant challenges faced by managers and researchers. Austin and Villanova (1992) chronicle the long and somewhat sorry history of the "criterion problem" in personnel psychology. Although significant progress has been made in resolving some of the issues in the historic debate over how to best understand and measure job performance, there are still significant questions about the definition of the construct "job performance" and about the best methods for measuring the performance of individuals, teams, and organizations.

Correspondence concerning this article should be addressed to Kevin R. Murphy. E-mail: krm10@psu.edu

Address: Department of Psychology, The Pennsylvania State University, Moore Building, University Park, PA, 16802

Kevin R. Murphy, Department of Psychology, The Pennsylvania State University.

Although they represent the most common method for measuring job performance, supervisory performance ratings are regarded by researchers and practitioners alike as the Rodney Dangerfield of human resource management—they rarely get much respect (Bernardin & Beatty, 1984; Guion, 1998; Murphy & Cleveland, 1995; Wallace, 1974). Coen and Jenkins (2000) argued that performance appraisal should be banned entirely. Although not taking such a strong stand, Murphy and Cleveland (1995) outline the potential harm performance appraisals can cause in organizations and question whether the benefits of performance appraisal are likely to outweigh the costs. Osterman (2007) argued that validation research that depends on performance ratings as a main criterion measure might not produce trustworthy, interpretable results.

One reason for the continuing survival of performance appraisals is the lack of better

https://doi.org/10.1111/j.1754-9434.2008.00030.x Published online by Cambridge University Press

alternatives. Performance measures can be characterized as either objective (i.e., measures that require few judgments, such as production counts) or subjective (i.e., measures that rely on the evaluative judgment of fallible judges), and although in principle objective measures might be preferable, there is broad agreement that objective measures of job performance are not feasible in most settings. First, as Landy and Farr (1983) note, many objective measures have surprisingly low levels of reliability and show little consistency across what should be equivalent indices. This is most clearly illustrated by measures of absenteeism; Landy and Farr (1983) and Gaudet (1963) report over 40 different indices of absenteeism and note that the correlations between different indices are frequently near zero (Chadwick-Jones, Brown, Nicholson, & Sheppard [1971] review evidence of the limited reliability of absenteeism measures). Second, objective measures of output, sales, and the like tend to be available for only a limited number of jobs. For example, it would not be sensible to collect tardiness or absence measures from sales representatives or from corporate managers who may not have a predetermined or fixed workday. Third, and most important, objective measures of performance almost always exhibit criterion deficiency. That is, there are parts of virtually any job that might be represented well with production counts or other objective measures, but there are other aspects of job performance (e.g., teamwork, contextual performance) that are not easily amenable to objective measurement. For example, it is possible to count how many patients a physician sees during the course of a workday but that would probably not be an adequate measure of his or her performance. Finally, the use of objective measures of performance tends to skew performance management and reward systems toward the countable, which can have adverse effects of performance. For example, it is possible to evaluate police officers in terms of the number of tickets they write or the number of arrests they make, but performance management systems that focus entirely on countable aspects of performance will push individuals to focus on these aspects of the job and ignore other aspects of the job (e.g., developing and maintaining good relationships with the community) that are less amenable to counting.

Because of the shortcomings of objective performance measures, most systems for measuring job performance continue to depend at least in part, and often almost completely, on the evaluative judgments of supervisors or other stakeholders in organizations (Murphy & Cleveland, 1995). These subjective judgments about performance are most often collected in the form of performance ratings, in which a supervisor, peer, or other stakeholder is asked to evaluate the effectiveness of performance on a series of dimensions (Landy & Farr, 1980, 1983) and often to also make judgments about the overall performance and effectiveness of the individual over some fixed period of time (e.g., over the last year). Some performance assessment systems depend on rankings rather than ratings or combine assessments of several different aspects of performance into an overall performance score rather than requesting explicit judgments about overall performance levels, but the norm is to obtain judgments about both performance dimensions and overall performance from an employee's direct supervisor and to use those judgments as one basis for high-stakes decisions (e.g., promotions, raises; Murphy & Cleveland, 1991, 1995).

Limited progress in improving performance appraisals. One of the dominant themes of research on performance evaluations over the past 75 years has been the need to improve the quality of performance ratings, and in particular, to assess and strengthen links between job performance and ratings of job performance. Over this period, a number of rating scale formats have been proposed as ways of imposing structure on raters' judgments and making the task of performance assessment easier and more reliable; this line of research was especially active in the 1960s and 1970s. The most notable variation in this theme was the development of rating scales that

https://doi.org/10.1111/j.1754-9434.2008.00030.x Published online by Cambridge University Press

used behavioral anchors to illustrate the meaning of specific performance levels of various dimensions of performance (Jacobs, Kafry, & Zedeck, 1980). Alternate approaches required raters to make judgments about the frequency of specific behaviors (Latham & Wexley, 1977). Landy and Farr's (1980) review of performance appraisal research led them to call for a moratorium on scale format research based largely on evidence that modifications to rating scale formats had only small effects on the reliability, validity, and utility of performance ratings.

A second approach to improving performance ratings involved identifying specific types of rating errors (e.g., leniency error, halo error) and removing these through rater training (e.g., Bernardin & Walter, 1977) or statistical control (Landy, Vance, Barnes-Farrell, & Steele, 1980; Murphy, 1982). Unfortunately, efforts to isolate and reduce rater errors were ultimately not very successful, in part because of the complexities in determining when and where these errors occur and in part because of the limited relationship between rater errors and rating accuracy (Balzer & Sulsky, 1992; Cooper, 1981; Murphy & Balzer, 1989; Murphy, Jako, & Anhalt, 1993; Murphy & Reynolds, 1988).

A third strategy for improving performance ratings involves rater training. As noted above, training raters to avoid specific errors was not a notably successful strategy. Wegner, Schneider, Carter, and White (1987) show how training people to suppress particular thoughts can lead to a paradoxical increase in those exact thoughts. It is possible that training people to watch out for and avoid specific errors could similarly backfire. However, training is not always aimed solely at suppressing errors (Bernardin & Buckley, 1981; Bernardin & Walter, 1977; McIntyre, Smith, & Hassett, 1984), and it is possible to improve the quality of ratings with some forms of training (e.g., Noonan & Sulsky, 2001; Sulsky, Skarlicki, & Keown, 2002). However, even with improved training programs, performance ratings are often greeted with skepticism.

One of the latest approaches to approving performance appraisals is to use 360° evaluations, in which assessments are obtained from supervisors, subordinates, peers, and sometimes clients (Bracken, Timmreck, & Church, 2001). These systems are often designed for feedback rather than for obtaining assessments of performance that can be used to help make administrative decisions, but in principle, rating systems that pool information from many sources might overcome many of the weaknesses that continue to plague supervisory performance appraisals. However, as Murphy, Cleveland, and Mohler (2001) note, the Achilles Heel of these systems is the systematic tendency for raters who examine performance from different perspectives to give systematically different ratings. In general, agreement between raters is far from perfect (Viswesvaran, Ones, & Schmidt, 1996); agreement between raters who occupy different positions in organizations is even worse, leading to questions about the wisdom of pooling across perspectives (Murphy et al., 2001).

None of the strategies reviewed above have been completely successful in improving performance appraisal, leading some researchers to believe that they may be based on a faulty diagnosis of the problems that beset performance appraisal. In particular, several of these strategies are based on the implicit assumption that ratings do not adequately reflect job performance because raters lack the knowledge, tools, skills, opportunities to observe, and so forth, needed to measure ratee performance well. An alternative hypothesis is that raters lack the *motivation* to rate accurately (Banks & Murphy, 1985). A substantial body of research has emerged examining systematic distortions in ratings, focusing in particular on raters' decisions to provide performance appraisals that do not reflect their best judgment about the ratee's level of performance but that help advance some other goal being pursued by the rater (Bernardin & Beatty, 1984; Longenecker, Sims, & Gioia, 1987; Murphy & Cleveland, 1991, 1995). For example, Morin and Murphy (1999) examined systematic differences in evaluations of the same set of ratees when performance ratings were collected for different purposes

https://doi.org/10.1111/j.1754-9434.2008.00030.x Published online by Cambridge University Press

and under administrative conditions that either gave ratees access to ratings or kept ratings confidential.

Regardless of whether they emphasize the rater's capability or the rater's willingness to provide good measures of ratee performance, most reviews of performance appraisal research (e.g., Landy & Farr, 1980, 1983; Murphy & Cleveland, 1991, 1995) suggest that the relationship between job performance and ratings of job performance is likely to be weak or at best uncertain. Furthermore, it does not appear that efforts to improve performance appraisals have been particularly successful. It is not clear that there is a better alternative, but it does seem clear that the relationship between job performance and ratings of job performance is not likely to be strong.

Organizational strategies for improving performance appraisals. Efforts to improve performance appraisals have not been limited to researchers or consultants; many organizations have developed or implemented procedures designed to address particular problems encountered in performance appraisals. Two notable examples are forced distribution systems and group discussion and review systems. Welch (2001) advocates using forced distribution rating systems to identify the weakest performers in organizations; "rank and yank" performance appraisal systems focus on identifying and removing perennially weak performers. Forced distribution rating systems can help address rating inflation, but they are generally seen as less fair than absolute rating systems (Roch, Sturnburgh, & Caputo, 2007). These systems can pose motivational problems, especially for employees who perform very well in an absolute sense but less well in comparison with other members of their workgroups (McBriarty, 1988), and the effects of rank and yank on workgroup quality can be hard to sustain (Scullen, Bergey, & Aiman-Smith, 2005). Most fundamentally, forced distribution rating systems mask differences in performance across divisions and across workgroups. It is likely that some workgroups are more effective than others, but if forced distribution ratings are used, it is impossible to tell which groups perform well and which groups perform poorly.

Systems that incorporate performance discussion and review often require raters to compare, discuss, and justify their evaluations. For example, the manager of a division might require all of his or her subordinates to present and discuss the performance ratings they assigned to their subordinates. These systems are likely to help calibrate raters, and they are likely to discourage unrealistically lenient or harsh ratings, but they also make raters more vulnerable to social influence effects. That is, raters are likely to be under some pressure to conform with norms regarding rating distributions, the types of ratings assigned and feedback given, and so forth, regardless of whether their subordinates are performing well or poorly, which can reduce the accuracy and validity of ratings.

Like the strategies reviewed in the previous section, these organizationally based efforts to improve performance appraisal provide partial solutions to some common problems in performance appraisal, but they do not address the fundamental question of why raters seem to find performance appraisal so difficult. Different models of the relationship between job performance and ratings of job performance suggest substantial explanations for the difficulties raters seem to face and, more important, different avenues for improving performance appraisal.

Three General Models of the Relationship Between Job Performance and Performance Ratings

There are numerous models of performance rating and performance appraisal in organizations (e.g., DeCotiis & Petit, 1978; Landy & Farr, 1980; Murphy & Cleveland, 1995; Viswesvaran, Schmidt, & Ones, 2005; Wherry & Bartlett, 1982), each of which focuses on different variables, but these models usually take one of three general forms, illustrated in Figure 1.

https://doi.org/10.1111/j.1754-9434.2008.00030.x Published online by Cambridge University Press

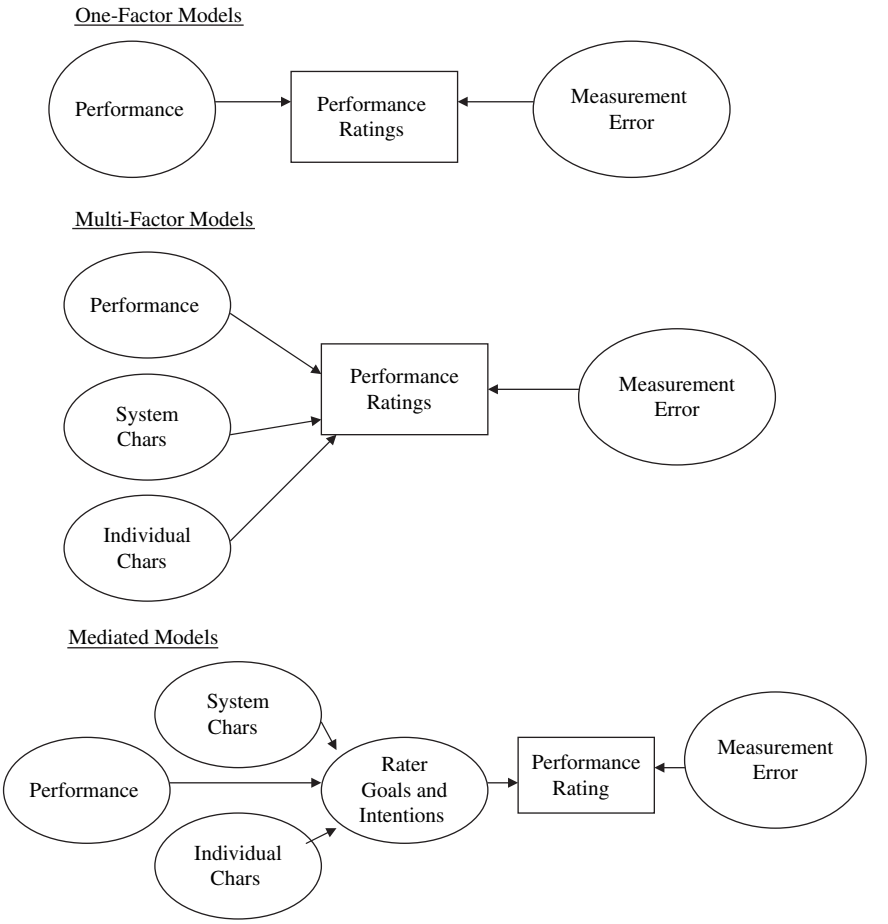

Figure 1. Three general models of the performance–performance rating relationship.

One-factor models. One-factor models suggest that the relationship between job performance and ratings of job performance is fairly straightforward and direct, but that this relationship is obscured by measurement error. This class of models does not require the assumption that performance is unidimensional, but it does argue that if measurement error is stripped away, there is a simple and direct relationship between job performance and performance ratings. For example, several papers by Schmidt and colleagues (e.g., Schmidt & Hunter, 1996, 1998; Schmidt, Viswesvaran, & Ones, 2000; Viswesvaran et al., 1996, 2005) have argued that there is a general factor present in most performance measures (objective and subjective) and that if the measurement error is taken into account, substantive conclusions about the relationships between various predictors and job performance can be made. Furthermore, proponents of one-factor models argue that strong inferences about the economic value of particular interventions (e.g., selection tests) can be made on the basis of correlations between scores on those tests and ratings of job performance (e.g., Hunter, 1983; Hunter & Schmidt, 1982; Schmidt & Hunter, 1998). For example, a recent paper by Le, Oh, Shaffer, and Schmidt (2007) describes in detail how to estimate the economic impact of using tests of general mental ability in personnel selections. The starting point for all these analyses is a meta-analytic estimate of the correlation between test scores and

https://doi.org/10.1111/j.1754-9434.2008.00030.x Published online by Cambridge University Press

measures of job performance, which in almost every case are supervisory ratings (Schmidt, 2002). After applying a variety of corrections for the imperfect reliability of performance ratings, range restriction validity studies, and so forth. Le et al. estimate the correlation between general mental ability and *job performance* to be .73, .66, and .39 in high-complexity, medium-complexity, and low-complexity jobs, respectively.

The distinguishing characteristic of one-factor models is that they partition the variance in performance ratings into two components, true score and error. Any researcher who is willing to make the leap from the observed correlation between test scores and performance ratings to the predicted increase in job performance and in the economic value of that performance must necessarily assume that the main difference between performance ratings and job performance is measurement error. Otherwise, correcting for attenuation, range restriction and the like would not be sufficient to allow researchers to say much about increases in job performance, solely on the basis of correlations between tests and performance ratings.

Proponents of one-factor models, citing several meta-analytic studies of interrater reliability estimates (e.g., Viswesvaran et al., 1996), often partition the variance in performance ratings roughly equally between true score and error components. In particular, because the average interrater reliability estimate reported in these meta-analyses is .52, it is common for researchers working within this framework to conclude that 52% of the variance in performance ratings reflects individual differences in job performance and that 48% of the variance in performance ratings reflects measurement error. A reliability estimate of .52 leads researchers working within this framework to interpret observed correlations between test scores and performance ratings as a consistent underestimate of the correlation between scores and job performance. Based on a reliability estimate of .52 for performance ratings, the corrected correlation between any variable and performance ratings will be nearly 1.5 times as large as the observed correlation. For example, if the observed correlation between test scores and ratings is .30, the correlation between test scores and job performance corrected for attenuation in the criterion is thought to be .41 or one over the square root of the reliability coefficient times the observed correlation.

The starting point for models that use interrater correlations to partition the variance in performance ratings into true score and error components is the entirely reasonable assumption that an individual supervisor is not likely to provide as good a measurement of performance as might be obtained by pooling ratings over several raters. Much in the same way that test scores become more reliable as more items are added to a test, performance ratings would arguably be more reliable and more valid if they were based on the pooled judgment of many raters. Murphy and DeShon's (2000a, 2000b) critique of one-factor models rests on three important differences between the classic psychometric model and the realities of performance appraisal. First, as Murphy and DeShon (2000a) note, it is rare to find multiple raters in organizations who have the opportunity to observe comparable aspects of ratee performance, making the notion of pooling across multiple raters logistically challenging. Second, agreements between raters cannot be directly linked to ratee performance and disagreements cannot be explained in terms of measurement error. Rather, there are systematic factors other than rater performance that can often lead two raters to agree in their performance ratings (Murphy & Cleveland, 1995), and there are factors quite distinct from classic measurement error that can lead them to disagree. Third, and most important, the interpretation of true scores that is implicit in one-factor models is problematic.

The true score component of performance ratings is usually interpreted in one-factor models as a measure of the job performance of the individuals rated (e.g., Le et al., 2007). As Lord and Novick (1968) have convincingly shown, the only useful interpretation of a true score is that it

https://doi.org/10.1111/j.1754-9434.2008.00030.x Published online by Cambridge University Press

represents an expected value over a sample of measurements. Because the causes of interrater agreement and disagreement are not limited to, and perhaps not even linked to ratee performance levels, there is no empirical or theoretical justification for interpreting this hypothetical true score as an indicator of job performance. Rather, the "true score" component of supervisory performance ratings represents a mix of systematic effects, some of which might lead raters to agree in their evaluations (e.g., the true performance levels of the ratees) and some of which might lead to systematic disagreements (e.g., differences in raters' willingness to give negative feedback).

One-factor models have two particularly striking characteristics. First, these models are unique in that they make no attempt to offer a substantive explanation for raters' apparent inability to measure ratee performance well. They simply accept this low reliability as a fact of life and go on from there. Second, they offer a seductively simple solution to the problems caused by the weak relationship between job performance and ratings of job performance. If this model is correct, many of the problems caused by the unusually low reliability of performance ratings can be solved through simple statistical corrections.

Because interrater correlations are in the neighborhood of .50, one-factor models assume that roughly 50% of the variance in performance ratings represents random measurement error that is completely idiosyncratic to the rater. More important, this error variance is assumed to be uncorrelated with true scores (which is true by definition; Lord & Novick, 1968) and, more important, uncorrelated with *any* other variable (e.g., test scores). This is the definition of measurement error that is part of the classical true score plus error model (Lord & Novick), but it has long been well known that this classic model is a special case of a much broader set of models of the generalizability of measures and that it is often a highly unrealistic one (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Lumsden, 1976; Murphy & DeShon, 2000a). Models that attempt to account for and understand interrater disagreements (including those discussed in the sections that follow) usually posit systematic as well as purely random sources of interrater disagreement and they lead to different conclusions about the meaning and value of performance ratings.

One-factor models may not offer practitioners clear solutions to the problem of improving performance appraisals, but they do offer a clear and simple solution to understanding the correlates and the meaning of these ratings. If you accept the logic of one-factor models, the solution to many of the problems caused by the poor performance of raters as measurement instruments is quite simple—all one needs to do is to correct for attenuation. It is important to note that correcting for attenuation does not make the supervisor's rating a better measure of an individual ratee's job performance. Corrected correlations can be useful for developing a better understanding of the correlates of performance ratings, but this correction is not intended or used to actually improve performance appraisals.

Murphy and DeShon (2000a, 2000b) regard one-factor models of performance ratings as untenable for several reasons. First, 80 years of research on performance appraisal suggests that breaking performance ratings into true scores and error components is not defensible (Murphy & DeShon, 2000a). If random measurement error is removed from performance ratings, the portion that is left is almost certainly not simply job performance but rather a composite of many systematic sources of variance, many of which have little to do with job performance (Landy & Farr, 1980; Murphy & Cleveland, 1995). In principle, true scores can be defined for any measure (e.g., the expected value of the rating obtained over a number of conditions is a type of true score—Lord & Novick, 1968), but these scores are likely to be very difficult to interpret; the sections that follow highlight these difficulties. Second, the 50–50 partition of the variance in ratings (i.e., the conclusion that about 50% of the variance in performance ratings is because of

https://doi.org/10.1111/j.1754-9434.2008.00030.x Published online by Cambridge University Press

random measurement error) is unrealistic (Murphy & DeShon, 2000a, 2000b). There is no good empirical or theoretical reason to believe that roughly half of the variance in performance ratings is because of random errors of measurement. Because there can be systematic factors that lead raters to disagree, it is possible for true performance to account for more than 50% of the variance in ratings of job performance and still see substantial levels of interrater disagreement.

This one-factor approach is based on a psychometric model of performance rating that is clearly unrealistic—raters in organizations do not function as parallel tests either in the strictest sense of the parallel test model (i.e., they do not produce scores with equal observed, true, and error variances) or in the less restrictive sense of weak true score models (Lord & Novick, 1968). Most important, this approach offers no explanation for the apparent inability of raters to report accurately which of their employees does his or her job well or poorly. The models that follow suggest that differences between job performance and ratings of job performance are not inexplicable, idiosyncratic errors on the part of raters, but rather are the result of a combination of systematic and random factors in the environment in which ratings are obtained. Rather than accepting the proposition that a large percentage of the variance in ratings in principle is inexplicable, multifactor models and mediated models attempt to develop and test substantive explanations for the apparent shortcomings of ratings as a measure of job performance.

Multifactor models. Landy and Farr (1980) developed an influential model of performance rating that delineated the potential influence of characteristics of jobs and organizations, rating scales, the purpose of ratings, characteristics of raters and ratees, and cognitive processes in evaluative judgment on performance rating. Their model is most often cited for its influence on a substantial body of laboratory research on cognitive processes in the 1980s and 1990s and for contributing to a substantial decline in research on rating scale formats over the next two decades (Guion, 1998, suggested that neither of these effects was entirely beneficial), but its most lasting contribution may be in laying out the range of variables other than job performance that are likely to influence performance ratings in organizations. Other multifactor models have been more narrowly focused; for example, DeCotiis and Petit (1978) focused on rater ability and rating norms; Wherry and Bartlett (1982) focused on decomposing bias in ratings; and DeNisi, Cafferty, and Meglino (1984) focused on cognitive processes. However, all these models share one of the key goals of the Landy and Farr (1980) model, providing a substantive explanation for the apparently poor performance of raters as judges of ratee performance.

Multifactor models suggest that the link between job performance and performance ratings is weak because many of the variables that systematically affect performance ratings (e.g., the rater's ability to recall relevant behavior, the purpose of ratings in organizations, the rating scales) have little or no relationship with the ratee's job performance. Thus, from the perspective of multifactor models, one important task faced by researchers and practitioners is to understand the nonperformance factors that affect performance ratings. Studies that attempt to isolate the effects of rater, ratee, and situational variables on performance ratings (e.g., Scullen, Mount, & Goff, 2000; Scullen, Mount, & Judge, 2003) have made considerable progress toward this goal.

Multifactor models suggest that the correlation between job performance and ratings of job performance depends largely on two factors: (a) the number of nonperformance factors that influence ratings and (b) the correlations between these factors and job performance. If there are a large number of systematic influences, most of which show only weak correlations with ratee performance, the effects of ratee performance on ratings of job performance will be masked by these other influences, and even if there is little random measurement error present in ratings, the expected correlations between performance and ratings

https://doi.org/10.1111/j.1754-9434.2008.00030.x Published online by Cambridge University Press

of performance will be small. In settings where there are fewer nonperformance influences, or where these influences are correlated with the ratee's performance (e.g., if raters find it easier to recall outstandingly good or outstandingly bad incidents, which will lead to larger correlations between performance and rating when there is substantial variability in true performance levels), correlations between performance and performance ratings will be larger.

Multifactor models suggest that the solution to the criterion problem is similar to solving the problem of a rattling noise in your car—that is, finding the potential sources of noise and dampening them. That is, this approach suggests that the key to improving performance appraisals is to identify the nonperformance factors that influence ratings and take steps to reduce their influence.

Mediated models. Like other multifactor models of performance rating, models presented by Murphy and Cleveland (1991, 1995; see also Cleveland & Murphy, 1992) attempt to identify factors in organizations and in performance rating systems that are distinct from ratee performance, but that nevertheless affect performance ratings. As with other models, these authors discuss the effects of rating source (see also Harris & Schaubroeck, 1988; Heneman, Wexley, & Moore, 1987; Scullen et al., 2000), the use and purpose of ratings (see also Cleveland, Murphy, & Williams, 1989; Jawahar & Williams, 1997; McIntyre et al., 1984; Murphy, Balzer, Kellam, & Armstrong, 1984), and contextual and attitudinal variables (see also Murphy, Cleveland, Kinney, Skattebo, Newman, & Sin, 2003; Tziner & Murphy, 1999; Tziner, Murphy, & Cleveland, 2001, 2002, 2005; Tziner, Murphy, Cleveland, Beaudin, & Marchand, 1998; Tziner, Murphy, Cleveland, Yavo, & Hayoon, in press; Williams & Levy, 1992).

The distinctive feature of Murphy and Cleveland's (1991, 1995) model is its emphasis on intentional distortions as an explanation for the weak link between job performance and performance ratings. Drawing on studies examining the influence of political factors on performance ratings (e.g., Bjerke, Cleveland, Morrison, & Wilson, 1987; Longenecker et al., 1987; Villanova & Bernardin, 1989), Murphy and Cleveland suggest that the goals being pursued by raters are critically important for understanding whether there will be a strong or a weak relationship between job performance and ratings of performance (see also Murphy, Cleveland, Skattebo, & Kinney, 2004).

Murphy and Cleveland (1995) suggest that a range of characteristics of individuals, organizations, and performance measurement systems lead raters to adopt different sets of goals when completing performance appraisals. For example, some raters use performance appraisals as a means of maintaining or increasing the performance of individual subordinates. A rater who believes that positive feedback will raise a subordinate's currently poor performance may choose to give that subordinate undeservedly high ratings. Other raters use performance appraisals to build and maintain positive interpersonal relationships with subordinates. Thus, a rater who believes that telling a poor-performing subordinate the truth about his or her performance will damage the ability of the workgroup to function might decide that it is better to inflate ratings than to provide accurate ratings.

Murphy and Cleveland (1995) also suggest that raters sometimes use performance appraisals to present a favorable image to upper management. They note that a large part of a manager's job is to get the best performance possible from his or her subordinates and that there are substantial disincentives to report that some or all subordinates are not performing well. Finally, they suggest that raters pursue internalized goals—that is, their own vision of what performance appraisals should be like. Thus, a supervisor who likes to see himself or herself as a tough, discerning boss might give lower ratings. A rater who believes that feedback about employees' strengths and weaknesses is more important that feedback about overall performance levels might emphasize within-person variability (i.e.,

https://doi.org/10.1111/j.1754-9434.2008.00030.x Published online by Cambridge University Press

patterns of strengths and weaknesses for each individual ratee) more than between-person variance (i.e., differences in overall performance levels).

Mediated models suggest that the solution to the criterion problem is to develop an understanding of the conditions under which raters will or will not attempt to convey their impressions of ratee performance when they complete performance appraisals. More than 40 years ago, Meyer, Kay, and French (1965) wrote about the difficulties caused by inconsistent roles in performance appraisal (e.g., supervisors are expected to be both coaches interested in developing subordinates and judges interested in evaluating them) and suggested that performance appraisals conducted for the purpose of giving employees performance feedback be separated from appraisals conducted for the purpose of making administrative decisions (e.g., raises, promotions). There is indeed evidence that the purpose of rating affects both the overall rating level and the correlates of performance ratings (Cleveland et al., 1989; Jawahar & Williams, 1997; McIntyre et al., 1984; Morin & Murphy, 1999; Murphy et al., 1984), but it is not clear whether separate appraisals would really work because they would tend to put raters in the position of saying different things about some ratees depending on the purpose. This inconsistency would likely undermine the defensibility of performance appraisal system.

This model suggests that performance ratings are often poor measures of job performance because raters are not willing to act as neutral measurement instruments. This also suggests a solution to the criterion problems. If organizations or researchers want to use performance appraisal ratings as measures of job performance, they need to structure the rating situation in such a way that raters have (a) incentives, tools, and opportunities to observe and recall ratees' job performance; (b) incentives to provide ratings that faithfully reflect the rater's evaluation of each ratee's performance; and (c) protection against the negative consequences of giving honest ratings. Most organizations have administrative handbooks that trumpet the importance and value of performance appraisal, but few if any reward raters for accurately evaluating their subordinates or sanction raters whose evaluations are systematically wrong (Murphy & Cleveland, 1995). If you want to use performance ratings as a measure of the performance of ratees, this model suggests that your main task will be to persuade raters to cooperate with this particular use of performance appraisals. Raters who believe that giving dishonest (usually inflated) performance ratings will help their subordinates improve, will improve the climate of the workgroup, and will make the rater look good to his or her superiors is not likely to give performance ratings that reflect the performance levels of his or her subordinates.

Conclusions

The three models presented here all agree on one essential point—that is, that the relationship between job performance and ratings of job performance is likely to be weak. They offer very different explanations for this weak relationship, ranging from random error to willful distortion, as well as different solutions for the criterion problem. Judging from the frequency with which interrater reliabilities are used in correcting validity coefficients for attenuation in meta-analyses and primary analyses, the one-factor model seems to be most popular, but it is also the model that stands up least well to close scrutiny (Murphy & DeShon, 2000a, 2000b). In particular, the belief that ratings of job performance reflect true performance and random measurement error is inconsistent with a quarter-century of research showing that factors other than ratee performance explain systematic variance in performance ratings. One-factor models also offer little in the way of explanations for or suggestions for improving raters' apparently limited ability to evaluate their subordinates accurately.

Multifactor models provide a useful starting point for improving performance appraisals. They suggest that the weak relationship between performance ratings and

https://doi.org/10.1111/j.1754-9434.2008.00030.x Published online by Cambridge University Press

performance is not likely to be entirely the fault of the rater but rather might reflect situational constraints and the influence of nonperformance correlates of ratings. Thus, this type of model is likely to lead to the conclusion that improving the rater (e.g., with training, better scales, behavior diaries, etc.) is only part of the solution and that systematic efforts to improve performance appraisal should also look at situational and contextual factors that interfere with the accurate measurement of performance.

Mediated models put the rater back at the center of the problem (and its solution), but they suggest a very different approach from that suggested by one-factor or multifactor models. Mediated models suggest that it is critical to engage raters as willing and motivated partners in the process of performance measurement and that doing so might require substantial changes in the way performance appraisals are used as a motivational tool, as a tool for maintaining group harmony, and as a self-promotion tool in organizations. Whether it will be possible or even advisable to use raters as performance measurement instruments remains to be seen (Murphy & Cleveland, 1995).

Academic debates about the causes of the weak relationship between job performance and ratings of job performance are interesting and useful, but they do not necessarily help practitioners solve the problem of improving performance appraisals. Both multifactor and mediated models do suggest some new avenues for improving the quality of rating data. Multifactor models place a premium on a careful analysis of the context within which performance is observed and ratings are obtained. These models suggest a two-track strategy for improving performance appraisals by increasing the rater's capacity to evaluate performance accurately (some strategies for increasing capacity, such as the use of behavior diaries and rater training, have shown some promise; Bernardin & Walter, 1977) and to remove the barriers to accurate assessment.

Mediated models put a premium on enlisting raters as an ally in the process of assessing performance in organizations. This set of models suggests that a useful first step is to work with organizations to determine whether they really want to measure performance and to make sure the costs and benefits of different ways of thinking about and using performance appraisals are clearly understood. The most obvious test of the proposition that an organization really cares about measuring performance accurately is to examine the sorts of rewards and sanctions that are applied to good- and poor-performance measurements. Raters are probably right to be skeptical about serving as neutral measurement instruments, and the practitioner's task in helping organizations develop high-quality performance measurement systems is more likely to be an exercise in organizational development than the standard exercise in scale construction.

If there is a real commitment to using performance appraisals primarily as performance measures, the practitioner's main job will be to give raters a combination of the tools and opportunities to measure performance well, the incentive to focus on this aspect of performance, and the protection from the negative consequences that can be associated with honest performance appraisals. Traditionally, practitioners have focused on building better scales or better training programs, but mediated models suggest that they might better focus their energies on building better climates for performance rating. These models suggest that the interventions most likely to improve the quality of performance appraisals in organizations are likely to look more like organizational development than scale development.

References

Austin, J. T., & Villanova, P. (1992). The criterion problem 1917–1992. *Journal of Applied Psychology, 77*, 836–874.

Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77*, 975–985.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research practice gap in performance appraisal. *Personnel Psychology, 38*, 335–345.

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work*. Boston: Kent.

https://doi.org/10.1111/j.1754-9434.2008.00030.x Published online by Cambridge University Press

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, *6*, 205–212.

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology*, *62*, 64–69.

Bjerke, D. G., Cleveland, J. N., Morrison, R. F., & Wilson, W. C. (1987). *Officer fitness report evaluation study* (Navy Personnel Research and Development Center Report, TR 88-4). San Diego, CA: NPRDC.

Bracken, D., Timmreck, C., & Church, A. (2001). *Handbook of multisource feedback*. San Francisco: Jossey-Bass.

Chadwick-Jones, J. K., Brown, C. A., Nicholson, N., & Sheppard, C. (1971). Absence measures: Their reliability and stability in an industrial setting. *Personnel Psychology*, *24*, 463–470.

Cleveland, J. N., & Murphy, K. R. (1992). Analyzing performance appraisal as goal-directed behavior. In G. Ferris & K. Rowland (Eds.), *Research in personnel and human resources management* (Vol. 10, pp. 121–185). Greenwich, CT: JAI Press.

Cleveland, J. N., Murphy, K. R., & Williams, R. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology*, *74*, 130–135.

Coen, T., & Jenkins, M. (2000). *Abolishing performance appraisals: Why they backfire and what to do instead*. New York: Berrett-Koehler.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

DeCotiis, T., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. *Academy of Management Review*, *3*, 635–646.

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

Gaudet, F. J. (1963). *Solving the problems of employee absence*. New York: American Management Association.

Guion, R. M. (1998). *Assessment, measurement and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Harris, M. H., & Schaubroeck, J. (1988). A meta-analysis of self-supervisory, self-peer, and peer-supervisory ratings. *Personnel Psychology*, *41*, 43–62.

Heneman, R. L., Wexley, K. N., & Moore, M. L. (1987). Performance-rating accuracy: A critical review. *Journal of Business Research*, *15*, 431–448.

Hunter, J. E. (1983). *The economic benefits of personnel selection using ability tests: A state of the art review including a detailed analysis of the dollar benefit of U.S. Employment Service placements and a critique of the low-cutoff method of test use* (USES Test Research Report No. 47). Washington, DC: U.S. Employment Service, USDOL.

Hunter, J. E., & Schmidt, F. L. (1982). Fitting people to jobs: Implications of personnel selection for national productivity. In E. A. Fleishman & M. D. Dunnette (Eds.), *Human performance and productivity. Volume I: Human capability assessment* (pp. 233–284). Hillsdale, NJ: Erlbaum.

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psychology*, *33*, 595–640.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, *50*, 905–926.

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

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. New York: Academic Press.

Landy, F. J., Vance, R. J., Barnes-Farrell, J. L., & Steele, J. W. (1980). Statistical control of halo error in performance ratings. *Journal of Applied Psychology*, *65*, 501–506.

Latham, G., & Wexley, K. (1977). Behavioral observation scales. *Journal of Applied Psychology*, *30*, 255–268.

Le, H., Oh, I., Shaffer, J., & Schmidt, F. (2007). Implications of methodological advances for the practice of personnel selection: How practitioners benefit from meta-analysis. *Academy of Management Perspectives*, *3*, 6–15.

Longenecker, C. O., Sims, H. P., & Gioia, D. A. (1987). Behind the mask: The politics of employee appraisal. *Academy of Management Executive*, *1*, 183–193.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology*, *27*, 251–280.

McBriarty, M. A. (1988). Performance appraisal: Some unintended consequences. *Public Personnel Management*, *17*, 421–434.

McIntyre, R. M., Smith, D., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, *69*, 147–156.

Meyer, H. H., Kay, E., & French, R. P. (1965). Split roles in performance appraisal. *Harvard Business Review*, *43*, 123–129.

Morin, D., & Murphy, K. R. (1999). Analyse empirique de la relation enre le contexte de l'évaluation de rendment et l'indulgence de l'évaluateur [The relationship between performance appraisal context and rating inflation]. *Relations Industrielles* [Industrial Relations], *54*, 694–726.

Murphy, K. R. (1982). Difficulties in the statistical control of halo. *Journal of Applied Psychology*, *67*, 161–164.

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

Murphy, K. R., Balzer, W., Kellam, K., & Armstrong, J. (1984). Effect of purpose of rating on accuracy in observing teacher behavior and evaluating teaching performance. *Journal of Educational Psychology*, *76*, 45–54.

Murphy, K. R., & Cleveland, J. N. (1991). *Performance appraisal: An organizational perspective*. Needham Heights, MA: Allyn and Bacon.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational and goal-based perspectives*. Thousand Oaks, CA: Sage.

Murphy, K. R., Cleveland, J. N., Kinney, T. B., Skattebo, A. L., Newman, D. A., & Sin, H. P. (2003). Unit climate, rater goals, and performance ratings in an

instructional setting. *Irish Journal of Management, 24*, 48–65.
Murphy, K. R., Cleveland, J. N., & Mohler, C. (2001). Reliability, validity and meaningfulness of multisource ratings. In D. Bracken, C. Timmreck, and A. Church (Eds.), *Handbook of multisource feedback* (pp. 130–148). San Francisco: Jossey-Bass.
Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology, 89*, 158–164.
Murphy, K. R., & DeShon, R. (2000a). Inter-rater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873–900.
Murphy, K. R., & DeShon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology, 53*, 913–924.
Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). The nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*, 218–225.
Murphy, K. R., & Reynolds, D. (1988). Does true halo affect observed halo? *Journal of Applied Psychology, 73*, 235–238.
Noonan, L. E., & Sulsky, L. M. (2001). Impact of Frame-of-Reference and Behavioral Observation Training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance, 14*, 3–26.
Osterman, P. (2007). Comment on Le, Oh, Shaffer and Schmidt. *Academy of Management Perspectives, 3*, 16–18.
Roch, S. G., Sturnburgh, A. M., & Caputo, P. M. (2007). Absolute vs. relative rating formats: Implications for fairness and organizational justice. *International Journal of Selection and Assessment, 15*, 302–316.
Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15*, 187–202.
Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.
Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.
Schmidt, F. E., Viswesvaran, C., & Ones, D. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*, 901–912.
Scullen, S. E., Bergey, P. K., & Aiman-Smith, L. (2005). Forced distribution rating systems and improvement of workforce potential: A baseline simulation. *Personnel Psychology, 58*, 1–32.
Scullen, S. E., Mount, M. K., & Goff, M. (2000). Evidence of the construct validity of developmental ratings of managerial performance. *Journal of Applied Psychology, 88*, 50–66.
Scullen, S. E., Mount, M. K., & Judge, T. A. (2003). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956–970.
Sulsky, L. M., Skarlicki, D. P., & Keown, J. L. (2002). Frame-of-reference training: Overcoming the effects of organizational citizenship behavior on performance rating accuracy. *Journal of Applied Social Psychology, 32*, 1224–1240.
Tziner, A., & Murphy, K. R. (1999). Additional evidence of attitudinal influences in performance appraisal. *Journal of Business and Psychology, 13*, 407–419.
Tziner, A., Murphy, K. R., & Cleveland, J. N. (2001). Relationships between attitudes toward organizations and performance appraisal systems and rating behavior. *International Journal of Selection and Assessment, 9*, 226–239.
Tziner, A., Murphy, K. R., & Cleveland, J. N. (2002). Does conscientiousness moderate the relationship between attitudes and beliefs regarding performance appraisal and rating behavior? *International Journal of Selection and Assessment, 10*, 218–224.
Tziner, A., Murphy, K. R., & Cleveland, J. N. (2005). Contextual and rater factors affecting rating behavior. *Group and Organizational Management, 30*, 89–98.
Tziner, A., Murphy, K. R., Cleveland, J. N., Beaudin, G., & Marchand, S. (1998). Impact of rater beliefs regarding performance appraisal and its organizational contexts on appraisal quality. *Journal of Business and Psychology, 12*, 457–467.
Tziner, A., Murphy, K. R., Cleveland, J. N., Yavo, A., & Hayoon, E. (in press). A new old question: Do contextual factors relate to rating behavior?—An investigation with peer evaluations. *International Journal of Selection and Assessment.*
Villanova, P., & Bernardin, H. J. (1989). Impression management in the context of performance appraisal. In R. A. Giacalone & P. Rosenfeld (Eds.), *Impression management in the organization* (pp. 299–314). Hillsdale, NJ: Erlbaum.
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.
Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108–131.
Wallace, S. R. (1974). How high the validity? *Personnel Psychology, 27*, 397–407.
Wegner, D. M., Schneider, D. J., Carter, S. R., III, & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology, 53*, 636–647.
Welch, J. F. (2001). *Jack: Straight from the gut.* New York: Warner Books.
Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35*, 521–555.
Williams, J. R., & Levy, P. E. (1992). The effects of perceived system knowledge on the agreement between self-ratings and supervisor ratings. *Personnel Psychology, 45*, 835–847.