CHAPTER 20

Job Performance Ratings

DANIEL A. NEWMAN, TED KINNEY, AND JAMES L. FARR

OVERVIEW OF CHAPTER 373
OVERVIEW OF PERFORMANCE RATINGS: A BRIEF HISTORY 373
BASIC CONSIDERATIONS IN THE RATING OF JOB PERFORMANCE 374
  Definition of Job Performance 375
  Rating Accuracy 375
  Jobs for Which Ratings Are Most Suitable 377
DEVELOPING AN APPLICATION OF PERFORMANCE RATINGS 378
  Clarifying the Purpose of the Rating System 378
  Operationalizing Performance 378
  Rating Format 379
  Rater Training 380

VALIDITY AND RELIABILITY OF PERFORMANCE RATINGS 381
  Validity 381
  Reliability 381
  More on Reliability in Performance Ratings 382
LIMITATIONS 382
CROSS-CULTURAL ISSUES IN PERFORMANCE RATINGS 382
LEGAL ISSUES IN PERFORMANCE RATINGS 383
FUTURE DEVELOPMENTS 385
  Rater Goals 385
  Rater Training and Cognitive Processes 385
  Rating Format Research 385
  Toward a Multilevel Conceptualization of Performance 386
REFERENCES 386

OVERVIEW OF CHAPTER

Job performance is a complex, multidimensional construct that can be defined and assessed in varying ways. Job performance can be defined (and assessed) in terms of quantifiable outcomes of work behaviors (e.g., amount of sales measured in dollars, number of academic journal publications, number of lines of computer code written) and in terms of behavioral dimensions (work-related communication, decision making, attention to detail) that are less quantifiable. It can be defined (and assessed) solely in terms of task performance (those activities that support the technical core of the organization and are a formal part of the relevant job description) but can also be defined as contextual performance (those activities that support the social and psychological environment of the organization and its employees; see Borman & Motowidlo, 1993).

Job performance can be assessed in terms of overall effectiveness in the job and in terms of specific components that the job comprises. Performance assessment information can be obtained from one or more individuals who are in one or more role relationships with the target employee (e.g., a supervisor, coworker, or subordinate).

In this chapter we focus on judgments (ratings) of the task performance of a target employee that are obtained from one or more individuals in a single role relationship with the target. Most commonly, the target employee is rated on several behavioral dimensions related to task performance and the source of the ratings is the employee's supervisor. Contextual performance ratings are discussed in Chapter 22 by Kaufman and Borman. Assessments that use ratings from individuals who are in several different role relationships with the target employee (frequently called multisource or 360-degree ratings) are discussed in Chapter 21 by Balzer, Greguras, and Raymark. We use the term performance ratings to refer to the judgments of an employee's task performance-related behaviors.

OVERVIEW OF PERFORMANCE RATINGS: A BRIEF HISTORY

The measurement of job performance is a fundamental assessment procedure in work organizations that serves as the basis for many personnel decisions and as the stimulus for more specific assessments of the individual employee. Job
performance ratings have been a primary assessment tool of
the industrial-organizational psychologist since the earliest
years of the field. Performance ratings were initially devel-
oped because of the recognition that judgments of
job performance were essential for those job functions for which
more objective measures were not available. Such ratings
were typically made by the supervisors of the group of em-
ployees whose job performance was evaluated.

Bingham and Freyd (1926) included two chapters on per-
formance rating scales in their book, Procedures in Em-
ployment Psychology. Several rating methods that had previously
been described in the research literature were discussed and
possible standards for evaluating the quality of the perfor-
ance ratings obtained by the use of rating scales were also
described. Among the possible indicators of rating quality
that Bingham and Freyd considered were interrater agree-
ment, the central tendency and dispersion of distributions of
ratings, intercorrelations among ratings of various job ele-
ments, and correlations between ratings of employees on spe-
cific aspects of the job and rankings of the same employees
on overall merit or value to the organization. Such indices of
the quality of a set of performance ratings would still be
reasonable ones today, suggesting the resilience of research
and application related to performance ratings in the first part
of the twentieth century.

Much of the focus of both research and application from
1925 to 1975 was devoted to a search for superior rating
instrument formats that would maximize the accuracy of the
resulting performance ratings or minimize what came to be
known as traditional rating errors. These errors included leni-
ency and severity (mean ratings that tend to be more fa-
vorable or more unfavorable, respectively, than warranted by
the individuals’ job performance), central tendency (almost
all ratings in the middle of the scale and almost none at the
endpoints), and halo (tendency of ratings of various dimen-
sions of job performance to be more highly correlated than
warranted by individual's job performance).

The basic assumption of the rating error approach was that
the measurement of employee job performance could be con-
ceptualized in a manner similar to the way that classical test
theory modeled the measurement of human abilities. Thus,
rated job performance was viewed as a function of "true job
performance," along with systematic and random error. The
accuracy of rated job performance would increase as error
decreased. Rating instrument format was hypothesized to have
an important effect on systematic rating error and received
much research attention. However, the results of rating in-
strument format research were generally disappointing, and
Landy and Farr (1980) called for a moratorium on research
on rating format, following their extensive review of the rat-
ing research literature.

Landy and Farr (1980) also suggested that research atten-
tion be given to the cognitive processes involved in perfor-
man ce rating. Their suggestion was consistent with a theory
of rating that Wherry proposed in the early 1950s but that
was not well known until its publication both as a journal
article (Wherry & Bartlett, 1982) and as an appendix in a
performance measurement volume (Wherry, 1983, in Landy
& Farr, 1983) about 30 years later. Several other cognitive-
based models of the rating process were also suggested in the
1980s, including those of Ilgen and Feldman (1983) and
DeNisi, Cafferty, and Meglino (1984). Rating research in the
1980s and 1990s did shift to an examination of cognitive fac-
tors affecting the processing of performance information by
raters and the resulting ratings of that performance. Cognitive-
based theories and research have increased our knowledge
about how such processes as attention, categorization of in-
formation, encoding and storage, and memory and recall in-
fluence the judgments that raters make. However, concern
has been expressed that the impact of the cognitive approach
on actual ratings of job performance in work organizations
has been minimal to date (e.g., Murphy & Cleveland, 1995).
One advantage of cognitive research and theory has been its
important contribution to the development of effective rater
training, a point that is addressed in more detail later.

Murphy and Cleveland (1995) emphasized the role of con-
textual factors on the rating process, suggesting that a shift
was needed from research and theory about performance rat-
ing judgments to research and theory about performance ap-
raisal. They noted that much of the cognitive-based research
was concerned with rating judgments that were never con-
veyed to the ratees. Performance appraisal implies that the
raters' judgments are communicated to the ratees as part of
a performance feedback process. A number of contextual
variables affect performance appraisal, including rater goals,
organizational norms, culture, and politics, and purpose of
the rating process (Murphy & Cleveland, 1995). An impor-
tant implication of this perspective is that the effectiveness
of a performance rating and appraisal process is also depen-
dent on these contextual factors.

BASIC CONSIDERATIONS IN THE RATING OF
JOB PERFORMANCE

Rating job performance accurately requires that both job per-
formance and rating accuracy be defined in a manner appro-
priate to the type of job being rated.
Definition of Job Performance

Before setting about the task of operationalizing and assessing job performance, it is critical to have an adequate conceptualization of what “performance” is. Without an appropriate specification of what does or does not constitute successful job performance in a given context, the usefulness of performance ratings will be unknown.

Historically, ratings of an individual’s job performance have included judgments about such diverse content as job behaviors (e.g., interpersonal interactions with a customer), job outcomes or results (e.g., dollar value of products sold by the employee), and individual traits (e.g., sociability). Although behaviors, outcomes, and traits are all intuitively germane to the performance concept, they are not all equally appropriate as the content of rating judgments, as will be discussed later. Summarily, behavioral measures of job performance have been acclaimed both as a more accurate basis than results measures for indexing individual job success (see Campbell, Gasser, & Oswald, 1996; Campbell, McClory, Oppier, & Sager, 1993; Ouchi & Maguire, 1975), and as easier to define and observe than trait measures (Wexley & Klimoski, 1984). We echo Campbell et al. (1993) in recommending that job performance ratings be limited to judgments concerning observable employee behaviors.

When studying job behaviors, researchers have classically thought of job performance as a unidimensional construct, citing intercorrelations among various performance measures as their primary evidence (Viswesvaran, 1993). In contrast, recent attempts to specify the types of behaviors that together constitute the job performance domain have produced some overlapping multidimensional conceptualizations of performance, which vary in terms of both content and specificity (e.g., Borman & Brush [1993] review managerial performance; Borman & Motowidlo [1993] review prosocial contextual performance). One of the more influential multifactor models of job performance is one advanced by Campbell et al. (1993), which specifies eight general, interrelated aspects of performance: (1) job-specific task proficiency, (2) non-job-specific task proficiency, (3) written and oral communication task proficiency, (4) demonstration of effort, (5) maintenance of personal discipline, (6) facilitation of peer and team performance, (7) supervision/leadership, and (8) management/administration. Although the importance of these eight aspects of performance will differ across specific jobs, they provide a widely applicable set of job behaviors that constitute a useful starting point for defining job performance.

Although performance ratings are commonly based on the rater’s judgment of the target’s average or “typical” performance, some performance appraisal scholars have noted that performance can be a dynamic criterion, exhibiting intrindividual variability over time. Kane (1986) has suggested that the ubiquitous use of “typical” performance ratings should be supplemented by raters’ assessments of the target employee’s entire performance distribution, based on the rater’s judgments of the frequencies of performance behaviors along a continuum from least effective to most effective. Distributional assessment has two proposed advantages: it has the conceptual advantage of incorporating the notion of consistency into the performance criterion, and it has the cognitive advantage of being more resistant to the memory biases and distortions usually found in judgments of typical performance, which require “mental averaging” prior to rating (see Kane, 2000). Building on the notion of distributional assessment, Sackett and colleagues have emphasized the distinction between an individual’s maximum performance and typical performance, demonstrating that maximum and typical measures are only weakly related (Sackett, Zedeck, & Fogli, 1988), with maximum performance showing the closer relation to cognitive ability measures (DuBois, Sackett, Zedeck, & Fogli, 1993). Last, research has demonstrated that there can be reliable individual differences in performance trajectories across time (e.g., Hofmann, Jacobs, & Baratta, 1993; Ployhart & Hakel, 1998).

To summarize, when using job performance ratings as an assessment of an individual employee, more information about that employee is obtained when the rating procedure contains judgments about multiple aspects of job behavior, about the individual’s typical and maximum job behavior, and about the frequencies of more and less effective behaviors. The specific format of the final rating instrument does not generally have a major impact on the quality of the ratings that are obtained, given that the behavioral dimensions and scale anchors are developed from a careful analysis of the job and its context.

Rating Accuracy

Performance ratings obtained for an assessment of an employee may be assumed to be accurate without explicit consideration of the factors that influence rating accuracy. This section discusses what is meant when the term accuracy is used, how accuracy has been measured, and several important conditions that can lead to accurate ratings. Finally, we discuss whether rater accuracy is always the most important goal when developing a performance assessment system.

The term rating accuracy has traditionally referred to the degree to which the judgment made by a rater represents the actual performance of the target (i.e., the target’s true score). One weakness of such a definition is that the true score of
a ratee is an unknown, making it impossible to determine whether the rater is accurately judging performance. In a multiple-rater situation, accuracy is often mistakenly considered to be the extent to which multiple raters' evaluations agree. That is, when there are multiple raters, interrater reliability or consensus is frequently considered a proxy for accuracy. In fact, it is possible for the raters to agree, but for the ratings to not represent the target's true score (Kenny, 1991; Murphy & DeShon, 2000).

When each target performer is rated by only one rater, it is not possible to compute the interrater agreement index. Researchers have developed several other accuracy measures that can be applied in such situations, where there are a number of employees, each of whom has been rated. Different accuracy indices vary in their level of complexity, but each operationalization of accuracy attempts to accomplish the same goal, which is to provide a 'true-score' that approximates the rating given by a hypothetical perfect rater under ideal circumstances (Murphy & Cleveland, 1995).

Typically, accuracy measures fall into two categories: evaluative accuracy and observational accuracy (Cardy & Dobbins, 1994; Cardy & Krzywinski, 1998). Evaluative accuracy measures indicate the accuracy of the rater's evaluations of performance level. This type of measure is tabulated as the squared difference between the actual rating and a true score. Observational accuracy measures indicate whether or not the rater identified the occurrence of various ratee behaviors (Lord, 1985). These types of measures reflect the accuracy with which the rater recalls or recognizes behaviors that were displayed by the ratee. For details of these measures, see Murphy and Cleveland (1995). It is important to note that the various measures of evaluative accuracy are not highly correlated, so that if a single such measure is used to determine accuracy, only part of the accuracy picture will be uncovered (Cardy & Dobbins, 1986). Generally, multiple measures of accuracy should be used (Murphy & Cleveland, 1995).

Murphy (1991) suggested that no one type of accuracy measure is preferable over another, although some researchers advocate one or the other type (e.g., Padgett & Ijig, 1989, prefer observational methods). Further, the relative applicability of an accuracy measure may depend on the research question or particular assessment purpose. For example, Lord (1985) suggests that observational accuracy measures are more important when researching rater memory. Additionally, Murphy (1991) suggested that observational accuracy may be better suited when the goal is to provide behavioral feedback or to justify an administrative decision. On the other hand, Murphy (1991) pointed out that if the goal is to understand the quality of the judgments made by a rater, then evaluative indexes of accuracy may be more important. The bottom line is that there is not one accuracy measure that is always preferred.

Research and practice do suggest that some general conditions can lead to more accurate ratings of job performance, regardless of the accuracy measure chosen. An important condition to ensure accurate ratings is whether the rater has an adequate opportunity to observe the ratee's behavior (Borman, 1978; Kenny, 1991). Kenny (1991) suggested that the more information a rater has about an employee, the more accurate the rater's evaluation will be. The point is that the more opportunities the rater has to observe behavioral samples of the ratee, the more likely the rater will be able to recall and accurately use those behaviors when making judgments. Rothstein (1990) found empirical support for the importance of the rater's observational opportunity for the accuracy of performance ratings in a work setting.

Another necessary condition for accurate ratings concerns the rater's familiarity with the rating format. As discussed in a later section of this chapter, performance rating instruments can take many forms. A rater must be familiar with and have practice using the chosen form. This consideration should not be taken lightly. For example, if a rater is accustomed to making behavioral ratings, then a trait-rating format may be difficult for the rater to understand. Further, to encode the appropriate information, raters should be familiar with the type of information that they will be asked to recall (Murphy & Cleveland, 1995). Cheung (1999) listed this source of inaccuracy as a common source of error variance in performance ratings. To prevent this problem, raters should have adequate opportunities to practice using the rating instrument.

Related to familiarity with the rating format is familiarity with the performance domain. Ratings can only be accurate to the extent that the rater is accustomed to the performance domain. Feldman (1985) suggested that raters should have experience with the rating domain based on performing or supervising in the performance domain, and raters should have training and experience in rating the performance domain. To meet this condition it is crucial that the performance domain be well defined and clearly communicated to the rater. For example, if the rater is under a false impression that a global rating of job performance will be asked for, but the ratings are separated into several dimensions, the rater may have difficulty recalling behaviors unique to each dimension.

Further, the rater must have an in-depth understanding of what types of behaviors make up each of the performance domains. Cheung (1999) suggested that when different raters have different perceptions of the composition of a performance domain, the evaluations the raters make may not be reliable. For example, if a rater is asked to assess a target's communications skills, then the organization must be explicit.
about what type of communication is relevant. If the organization is interested in both written and verbal communication, but the rater only considers written communication in his or her assessment, accuracy will likely suffer.

Another condition that has received research attention is accountability (see Antonioni, 1994; Klimoski & Inks, 1990). Accountability refers to the organizational or social pressure for a rater to justify his or her ratings (Tetlock, 1992). Although it was initially hypothesized that a rater who is held accountable for ratings will be more likely to provide accurate ratings, a frequent finding is that raters distort ratings in a positive direction if they are accountable (Antonioni, 1994). This finding has led many researchers to consider accountability as a source of bias in ratings (Stone, Rabinowitz, & Spool, 1977). However, more recently Mero and Motowidlo (1995) found that raters who were accountable made more accurate ratings than anonymous raters. Thompson (1995) suggested that holding raters accountable may increase accuracy if the raters do not have a goal to achieve a predetermined conclusion, but that accuracy may suffer when they do. Haeggborg and Chen (1999) found that the effect of accountability depended on the accuracy measure that was used. No unequivocal conclusions can yet be reached about the effect of accountability on rating accuracy. Thus, the question of whether to hold raters accountable for ratings during a specific application should be given careful consideration in each situation.

A final contextual factor relevant to rating accuracy is the motivation of the raters. Although it is frequently assumed that raters will always choose to provide the most accurate ratings possible, this assumption may be untenable (Murphy & Cleveland, 1991, 1995; Tetlock, 1992). Murphy and Cleveland (1991, 1995) have discussed how rater goals influence performance ratings. For example, if a rater decides that an employee deserves to be promoted, the rater may give that employee ratings that will achieve the desired outcome as opposed to giving the most accurate ratings.

To combat a rater’s goals to rate in a manner that is not accurate, some researchers have suggested that rewards be offered for accurate ratings (see Mohrman & Lawler, 1983). The problem with this idea is that even if rewards are offered for accurate rating, such accuracy can be difficult to measure and, thus, the link between accurate ratings and subsequent rewards may be weak. If a rater does provide accurate ratings and is not rewarded, the rater may then begin to distrust the performance appraisal system (Murphy & Cleveland, 1995) and start a negative spiral of reduced rating accuracy over time. The negative consequences surrounding the giving of low ratings may also outweigh the rewards. Therefore, rewarding accurate rating does not seem to be an effective means to boost accuracy in ratings. One way to combat divergent rater motivation is to address and stress the importance of accurate rating in a training program. Although it is doubtful that this solution will remedy the problem entirely, it appears to be one of the more feasible options.

A final point is whether it is appropriate to consider rating accuracy the sole legitimate goal of the users of a performance assessment system. Murphy and Cleveland (1995) argued that accurate ratings are not always a positive outcome for all users. These authors suggested that the various indexes of accuracy are context-free, but performance ratings are not given in a context-free environment. Many organizational forces influence ratings. Murphy and Cleveland (1995) pointed out that ratings are a means of communication by the rater to the organization and its members. It is sometimes possible that the most crucial message is not communicated with strictly accurate ratings. That is, inaccurate ratings may deliver a more meaningful and important message to organizational members than accurate ratings. For example, if a rater inflates the ratings of an employee whom the rater believes should be promoted, then rating inaccurately could lead more efficiently to the appropriate employee promotions than accurate ratings could.

To summarize, while rating accuracy is a somewhat controversial topic, some clear-cut recommendations are possible. First, although existing rating accuracy measures are by no means perfect indicators, they do impart some useful information about the quality of a set of performance ratings. Second, some conditions lead to more accurate ratings. The raters should always have ample opportunities to observe the target’s performance. Also, the raters should be intimately familiar with the rating format and the performance domain. Finally, other issues such as accountability and the importance of accurate ratings should be considered.

Jobs for Which Ratings Are Most Suitable

Performance ratings can be applied to almost any job and its various performance domains. The exceptions would be those relatively rare domains of specific jobs for which an accurate and bias-free objective or countable measure does exist. An assessment should include these results measures only when they are judged by relevant organizational members as providing useful and valid information, given the purpose for which performance information is to be used.

The exact performance domains and the specific form of the rating instrument depend on the information provided from an analysis of the job in question. Details of these procedures are provided next.
DEVELOPING AN APPLICATION OF PERFORMANCE RATINGS

When developing a performance rating system for use in organizations, five important choices must be made. These choices are illustrated by the following questions:

1. What is the purpose of the rating system?
2. What perceptions, behaviors, outcomes, and traits constitute an appropriate operationalization of performance for this purpose?
3. What is the appropriate format for collecting rating information?
4. Who should make the ratings?
5. How should the raters be trained prior to making ratings?

All five of these questions are interdependent, and each of the five choices they represent is indispensable to the overall rating process. Next we offer a brief discussion of alternative answers to the five questions.

Clarifying the Purpose of the Rating System

In 1989, Cleveland, Murphy, and Williams reported results from a large-scale survey describing the uses to which performance ratings are put. Four general uses of performance appraisal were identified and labeled: (1) between-individuals comparisons (salary administration, promotion, retention and termination, individual recognition, layoffs, identifying poor performers), (2) within-individuals comparisons (identifying individual training needs, performance feedback, determining transfers and assignments, identifying strengths and weaknesses), (3) systems maintenance (personnel planning, determining organizational training needs, evaluating goal achievement, assisting in goal identification, evaluating personnel systems, reinforcing authority structure, identifying organizational development needs), and (4) documentation (documenting personnel decisions, defining criteria for validation research, and meeting legal requirements). Of these four purposes, performance ratings were most often used for between- and within-individuals comparisons. Interestingly, most organizations used performance ratings for multiple purposes simultaneously.

The purposes an organization has for ratings can influence the ratings given. Research has shown that rating behavior varies as a function of the purpose for which the rater believes the ratings will be used (Jawahar & Williams, 1997). With more lenient ratings given when salaries and promotions are on the line, compared with ratings made for individual feedback or research purposes only. Furthermore, organizational purposes are not the only purposes relevant to rating; individual raters can have their own agendas for the ratings, such as protecting a subordinate from disciplinary action or avoiding negative attention that would be brought to a work unit with low-rated individuals (Longenecker, Sims, & Gioia, 1987).

Operationalizing Performance

Closely related to the choice of purposes for a performance rating system is the decision of what to measure in order to achieve those purposes. Many options are available when attempting to quantify performance, although the multiplicity of options is often blurred by the blanket use of the term performance to refer to any and all of them. The basic options for operationalizing performance have been reviewed by many performance rating researchers (Curd & Dobbins, 1994; Landy & Farr, 1983; Smith, 1976; Wexley & Klimoski, 1984), and they are reviewed again here.

Four common operationalizations of performance are:
1. Subjective criteria (e.g., ratings of overall performance),
2. outcomes/results/nonjudgmental criteria (e.g., sales, turnover, product quality, speed),
3. Behavioral criteria (e.g., duties and tasks), and
4. Trait criteria (e.g., knowledge, skill, experience, motivation).

Each of the operationalizations has strengths and limitations. Subjective criteria suffer from vulnerability to bias and error, can be manipulated for political purposes, or to advance an individual rater's personal agendas, and may be perceived by the ratee as invalid or unfair, particularly when negative feedback is given. Results criteria have the disadvantage of not being completely under individual control. Because the quality or amount of product someone makes can depend in large part on the initial material inputs, resources, and opportunities available for obtaining positive results, these criteria are often more indicative of system performance than individual performance. It can be perceived as unfair to discipline or reward individuals on the basis of results "performance" over which they had little personal control (e.g., low total sales of luxury items during an economic recession). Advantages of results criteria are that they are closely linked to financial indexes of organizational productivity/effectiveness and are "objective" in the sense that they are less subject to rating errors. Behavioral criteria provide a partial solution to the problem of person versus situation attributions for performance, since behaviors are more under individual control than results criteria. Behavioral criteria are also a good basis for feedback. Last, trait criteria can suffer from rating bias and error and are only useful when it is possible to move personnel to different positions or teams.
In any attempt to select operational criteria for a performance management system, one should keep in mind that the intended purpose of the performance ratings is itself the ultimate criterion. If ratings are being made for the purpose of promotion decisions, then the ultimate criterion is the relative proficiency of those who are promoted over those who are not. If ratings are being made for the purpose of individual development, then personal change is the ultimate criterion. Specifying the criterion for a performance rating system is a problem of measurement validity and therefore merits a careful consideration of criterion contamination and deficiency, as well as convergent, discriminant, and substantive construct validity (Cronbach & Meehl, 1955). Meta-analysis has shown the correlation between objective and subjective performance measures to be low, indicating that these two types of measures are reflections of somewhat distinct concepts (i.e., have low convergent validity) (Bommer, Johnson, Rich, Podsakoff, & Mackenzie, 1995).

The performance construct has two important features that have been frequently overlooked by past research: (1) performance time frame (short- vs. long-term; Rambo, Chomiak, & Rountree, 1987), and (2) unit of analysis (individual performance vs. team performance vs. organizational performance; DeNisi, 2000; Schneider & Klein, 1994). It is expected that future research on the performance construct will elaborate these temporal and multilevel features, particularly in areas where the intended purposes of performance rating emphasize the need for such operationalizations.

Rating Format

Historically, the choice of rating formats was believed to influence rating accuracy and interrater reliability. The major rating format options have included graphic rating scales (GRS; in which an employee is rated on her exhibition of some trait or behavior using a Likert scale—e.g., from 1 to 5), behavioraly anchored rating scales (BARS; in which behavioral statements are developed to represent typical behavior at various levels of performance along the scale, helping the rater to calibrate the rating given; Smith & Kendall, 1963), behavior observation scales (BOS; in which raters judge the frequency with which each behavior was exhibited by the ratee; Latham & Wexley, 1981), mixed standard scales (MSS; in which behavioral statements are made for good, medium, and poor performance on each performance dimension, then the ratee's behavior is compared with each of the three performance-level statements for each aspect of performance on a three-point scale [i.e., more effective, equally effective, less effective]; Blanz & Ghiselli, 1972), behavioral summary scales (BSS; like BARS, but with more generic behavioral statements; Borman, 1979), performance-distribution assessment (which elicits both modal performance and variance in performance over time; Kane, 1986), and computerized adaptive rating scales (CARS; presents pairs of scaled behavioral statements iteratively, adapting presented pairs on the basis of previous responses; Borman et al., 2001). In a comparison of the various rating scales developed up to that point in time (e.g., BARS, GRS, etc.), Landy and Farr (1980) demonstrated that various rating formats were virtually indistinguishable in their effects on rater error, and they concluded that the choice of format matters little as long as ratings are behavior based and job related.

Selecting a Rating Source: Who Should Make the Ratings?

Performance ratings can be made by several different sources, including supervisors, peers, subordinates, external customers, skip-level supervisors, and even self-ratings. Each of these sources is likely to be privy to unique types of information, to have some degree of shared perspective based on the source's position relative to the rating target, and to have personal agendas for the ratings. For instance, supervisors are likely to be knowledgeable about the target's productivity, peers are likely to be knowledgeable about the opportunities and constraints involved in a particular job, and external customers are likely to be knowledgeable about the quality of the target's work product.

When selecting a rating source, one should think about how the ratings will be used. If the ratings are part of a promotion decision, then asking for ratings from a peer who is in competition for the promotion can produce a conflict of interest. If the purpose of the ratings is to give the ratee developmental feedback but with no consequences for organizational reward or punishment, then there might be advantages to letting ratees select their own raters from among those who are likely to be knowledgeable.

A great deal of recent research and practice has incorporated the idea that multiple sources should be used to rate each target (see Chapter 21 of this volume; Bracken, Timmerman, & Church, 2001). Such multirater systems have special implications for rater training, choice of rating instrument, and rating purpose. With regard to the selection of rating sources for a given application, Farr and Newman (2001) have noted that little empirical research exists to guide source selection. Until such research accumulates, prescriptions for source selection must be based on common sense and the general finding that rating reliability improves as raters have more opportunity to observe ratees (Rothstein, 1990).
Rater Training

It is important to remember that in the performance rating situation, the raters are the measurement instrument because raters are subject to many idiosyncratic social, political, and cognitive sources of variation, it is naive to assume that all raters give equally valid ratings by default. After the rating source is identified, the raters must be trained. Without training, the different measurement instruments (i.e., raters) cannot be expected to make precise assessments. To accomplish this task several types of training programs have been developed and implemented. These training programs include rater error training (RET), frame of reference training (FORT), performance dimension training (PDT), rater variability training (RVT), and behavioral observation training (BOT). Some of these training programs have been around for 30 years (e.g., RET), while others are newer (e.g., FORT), but all of them attempt to improve ratings. Evidence regarding the effectiveness of each of these programs also is discussed.

Rater error training has been used in performance rating situations for several decades. Its general purpose is to improve rater accuracy by minimizing rater errors, discussed earlier in this chapter, such as central tendency, leniency, and halo. This approach is probably still the most common form of rater training. The problem is that RET does not seem to achieve its purpose very well. Several sources have reported that accuracy drops or fails to improve after RET programs have been implemented (Bernardin, Cooke, & Villanova, 2000; Hauenstein, 1998; Woehr & Huffcutt, 1994). The consensus seems to be that RET is not particularly effective but is frequently implemented.

Among rater training researchers, FORT seems to have taken the place of RET as the dominant type of training. The main problem with FORT is that most uses of FORT have been in laboratory settings (Bernardin et al., 2000), with relatively little evidence accumulated for the effectiveness of FORT in applied settings. Keown-Gerrard and Sulsky (2001) stated that FORT focuses on calibrating raters so they agree on the dimensions on which performance is judged and what constitutes different levels of performance. Bernardin and Buckley (1981) noted that FORT emphasizes the multidimensionality of performance, the definition of performance, defining and describing performance examples, and practice and feedback using these established standards.

The steps in developing a FORT program are straightforward but require several judgment calls during the development of the procedure. For example, one of the goals of a FORT program is to help the raters understand the definition of performance within a given setting, which creates the need for training the program developers to make specific decisions about what will and will not be included in the performance domain. The steps involved in creating a FORT program are as follows (Bernardin et al., 2000; Bernardin & Buckley, 1981):

1. A definition of performance must be generated. This definition must be very precise and should include descriptions of what types of behaviors should and should not be considered a part of the performance domain.
2. The raters must become familiar with performance dimensions and performance levels across dimensions. This task is accomplished by giving the raters the rating materials and meticulously going over every aspect of the rating format and the performance domain with the raters.
3. Presentation and discussion of critical incidents depicting different levels of performance across dimensions should be presented and discussed. This activity helps solidify the different dimensions of performance in the raters’ mind. Through this process the rater is exposed to clear examples of poor, average, and favorable behavioral examples of performance on the different performance dimensions.
4. The key step in FORT is to allow the raters to practice rating. The raters should practice rating using the rating instrument until they are entirely comfortable with its use and the performance domain being rated.
5. After the raters have completed assigning practice ratings, detailed qualitative and qualitative feedback sessions should be held with each rater. Through this process, the rater can gain a better understanding of how ratings should be assigned and how he or she can improve. If possible, these sessions should be followed by more practice and feedback.

Several reviews suggest that FORT is effective (Arvey & Murphy, 1998; Bernardin et al., 2000; Woehr & Huffcutt, 1994), using the criterion of rating accuracy. At the present time, FORT is in need of more research in applied settings.

Much less has been written about other types of training, including performance dimension training, rater variability training, and behavioral observation training. That said, Hauenstein (1998) conducted a thorough review of the existing work on these training programs. His conclusion was that these types of training can be effective, but not as effective as FORT. These results must be considered preliminary, however, because very few studies of rater training methods, other than RET and FORT, have been conducted (Bernardin et al., 2000).

Performance dimension training attempts to help the rater become more familiar with the rating system elements. PDT is not necessarily concerned with accuracy; rather, it helps the rater to discriminate between different dimensions of performance. Hauenstein (1998) stated that this training is nec-
essary but not sufficient for establishing rating accuracy. PDT could be thought of as a focus on steps two and three of a FORT program.

Rater variability training is designed to increase variability in ratings to reflect the variability in true performance. Although this practice may seem like a good idea, it may ultimately obtain similar results to those RET has received. The danger of this type of training is that it does not train raters how to rate a target’s performance but, rather, teaches raters how their rating distribution should look. This practice could be dangerous if the target group happens to have a true performance distribution other than the one prescribed by the training. Therefore, RVT may lead to rating inflation and deflation for the sake of attaining an attractive distribution.

Behavioral observation training improves detection, perception, observation, and recall of relevant behaviors. BOT methods such as keeping diaries have been shown to be effective.

The bottom line of this discussion is that there are several types of rater training to choose from, and the pros and cons of each type should be weighed when choosing any of them. FORT seems to be the highest regarded type of training in the research literature. PDT, RVT, and BOT seem to work, but not as well as FORT (although not enough research has been conducted on these types of training). RET, while still enjoying popularity, may be the least effective of the various types of training programs.

VALIDITY AND RELIABILITY OF PERFORMANCE RATINGS

When developing a performance rating application using the five aspects of method just outlined, a key goal is to produce ratings that are valid and reliable representations of an individual’s performance. Each of the five aspects of method—purpose, definition of performance, format, source selection, and rater training—has important implications for validity and reliability of the ratings.

Validity

Rating validity, or the extent to which the ratings measure what they are designed to measure, mandates clear decision making on the part of the rating system designer regarding what “performance” should mean for a particular application. If the purpose of the rating system is to change individuals’ performance (through feedback and development), then performance should be operationalized at the individual level in very specific, behavioral, and absolute terms. By contrast, if the purpose of assessing performance is to support decisions of which individuals should be moved (promotions, transfers, terminations), then performance can be operationalized as both behavioral and results criteria, measured relative to others in the work unit. Last, if the purpose of the performance appraisal is to aid in organizational redesign and strategy, then performance measures should include system-level results criteria that are aligned with the competitive strategy of the company (i.e., brokerage firms measure sales, grocery stores measure return customers, factories measure productivity and absences, pharmaceutical companies measure patents, etc.). Failure to align performance measures with the intended purpose of the appraisal system will likely result in construct validity problems, such as criterion contamination (measuring something in addition to what was intended) or criterion deficiency (not measuring all of what was intended) (Brogden & Taylor, 1950). The concept of measurement validity is inherently value-laden, and hence the goals of the rating system need explicit specification prior to selection of the rating instrument, raters, and rater training. The performance construct should be matched to the purpose for which ratings are used in content, level of specificity, and level of analysis (e.g., task vs. contextual performance, specific task performance vs. overall performance, team facilitation vs. individual performance vs. performance over time, etc.)

On the topic of rater source selection and rating instrument selection, one contemporary method for assessing the construct validity of performance ratings is to use a version of the Multitrait-Multimethod matrix (Campbell & Fiske, 1959). In this approach, ratings of the intended performance construct(s) can be made by multiple rating sources, using multiple instruments. Since each rating source is providing a rating of each performance dimension using each rating instrument, it is possible to statistically model the components of the ratings that are idiosyncratic to a specific rating source or instrument. It is also possible to establish the convergent validity of ratings across sources and instruments, as well as the discriminant validity of various performance criteria that are proposed to be theoretically distinct (for updated statistical techniques, see Conway, 1998; and Bagozzi, Yi, & Phillips, 1991). Conway (1999) provided an example of a multitrait-multisource approach for assessing the construct validity of managerial performance, in which he modeled the relative contributions of various performance facets (e.g., job dedication, interpersonal facilitation, task performance) to overall performance ratings made by both peers and supervisors.

Reliability

No rating can be valid without being reliable. The reliability, or consistency of measurement, of performance ratings has been assessed in several different ways. Although perfor-
Performance rating reliability has been indexed by both interrater reliability (among several raters) and intrarater reliability (across time or scale items). Schmidt and Hunter (1977) advised that the most appropriate measure of criterion reliability is reliability across two raters who rate the target at distant points in time. Viswesvaran, Ones, and Schmidt (1996) have critiqued the use of intrarater agreement as an index of reliability because it includes transient and idiosyncratic rater errors. These authors interpreted empirical research as favoring the use of interrater reliability and reported a meta-analytic estimate of interrater reliability of .52 for supervisor ratings of overall performance (leading to the conclusion that 25% of the variance in supervisor performance ratings is true score variance). In spite of these unimpressive results regarding the reliability of performance ratings, some researchers believe .52 to still be an overestimate of true reliability. Specifically, the use of intrarater reliability has been criticized because it includes systematic rating biases that raters share in common (Murphy & DeShon, 2000; but see Schmidt, Viswesvaran, and Ones [2000] for a critique of Murphy and DeShon's critique, positing that such views of measurement result in nihilism). Scullen, Mount, and Goff (2000) shed some additional light on the subject, using a generalizability theory framework (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) to demonstrate that over 50% of the variance in performance ratings is due to idiosyncratic rater effects (halo error), whereas only 25% of rating variance is due to the raters' actual performance. These results indicate that performance ratings may reveal more about the rater than about the ratee, and they provide a rich area for future research.

More on Reliability in Performance Ratings

The purpose of performance ratings may influence the ratings' apparent reliability as a result of distributional artifacts associated with rating purpose. Specifically, the finding that ratings made for between-persons decision making (e.g., promotions, terminations) are subject to leniency errors (Jawahar & Williams, 1997) implies that the range of such ratings would be restricted. This restriction of range (e.g., raters only using response options 6 and 7 on a 7-point scale) could result in an artificially inflated estimate of interrater agreement and could also artificially deflate estimates of interrater reliability based on the correlations among multiple raters' ratings. Obtaining estimates of the overall effect of rating purpose on rating reliability is an issue for future research.

The definition of performance also has implications for rating reliability, because some performance constructs (communication and interpersonal competence) are rated less reliably on average than other performance constructs (productivity and quality) (interrater reliability; Viswesvaran et al., 1996).

Using the measurement approach of item response theory (IRT; Hambleton, Swaminathan, & Rogers, 1991), Faust and Craig (2001) have demonstrated that raters from various rating sources (supervisors, peers, self) discriminate rates equally well across the range of performance. These results imply that reliability of performance ratings may not be greatly influenced by any systematic biases attributable to the rating source.

Last, rater training can be an important part of establishing rating reliability. Previously discussed techniques such as FORT can be implemented to improve the consistency of calibration between raters.

LIMITATIONS

The use of subjective ratings of individual performance as a basis for personnel decisions and feedback has many limitations. Such limitations are reflected throughout this chapter and pervade most discussions of the science and practice of performance appraisal. In particular, performance ratings are subject to the aforementioned systematic errors (e.g., leniency, halo, recency, contrast), can be manipulated in accord with rater goals (i.e., social and political motives), and depend on the rater's knowledge of and opportunity to observe job behavior. Furthermore, job requirements may vary across occasions and situations, or may be poorly defined, leaving the rater with the complex task of assessing a fuzzy, moving target. Last, performance ratings are rendered meaningless when put to purposes other than those understood by the rater (e.g., using ratings designed for feedback purposes as a basis for deciding whom to terminate). At the extreme, staunch critics have claimed that performance ratings are inherently antithetical to organizational productivity (e.g., see Deming, 1986).

CROSS-CULTURAL ISSUES IN PERFORMANCE RATINGS

Throughout the past 20 years culture researchers have suggested that culture can influence several human resource functions (cf., Erez, 1997; Hofstede, 1984; Triandis, 1994). Despite these suggestions, performance rating research involving culture has been slow to develop. Although culture has been assumed to affect performance rating, there is no consensus about the nature of culture's influence on rating and little empirical research from which to draw.

Evidence has been mounting that the cultural dimension of individualism-collectivism may be critical in understand-
ing performance rating. According to Bailey, Chen, and Duo (1997), individualism-collectivism refers to “cultural dispositions to understand oneself primarily in terms of either satisfying personal aspirations or attending to group needs.” Erez (1997, p. 606) pointed out that individual performance appraisal may not be acceptable in collectivist cultures because collectivists focus on the workgroup rather than the performance of any one person in the group. Oyserman, Coon, and Kemmelmeier (2002) pointed out several potential problems with the measurement of individualism and collectivism but still concluded that these are useful constructs. Space does not permit here a full discussion of the issues raised by Oyserman and colleagues, and the interested reader is referred to their paper.

As Triandis (1994) discussed, performance appraisals often assume that most of the variance in performance can be attributed to individuals (rather than group variance or situational variance). This assumption is based on research that has been conducted on performance appraisal in Western cultures. It is important to understand the effects that such an ethnocentric bias may have as organizations become increasingly more global. Triandis (1994) stated that more than half of the people in the world are socialized in collectivist cultures. If research in individualist cultures regarding performance ratings does not generalize to collectivist cultures, the use and interpretation of performance ratings as conceptualized in Western research traditions will be inappropriate for much of the world’s population.

It seems likely that differences in performance rating across the individualist-collectivist cultural dimension could be caused by differences in attribution strategy or by generic beliefs about the causes of behavior. Social psychological literature has discussed the role of attribution differences across this cultural dimension. According to Triandis (1996), understanding what causes things to occur in our world provides perceivers with some capability to make sense of—and to foresee the consequences of—actions and events. Several studies from the last 10 years can be summarized by the general conclusion that, as Triandis (1994) and Triandis and Bhawuk (1997) noted, collectivists attribute success externally (e.g., to luck, circumstance, etc.), whereas individualists attribute success internally (e.g., to effort). These differences in attribution suggest that raters from different cultures will perceive performance from a rater differently, and thus assign ratings differently. Kitayama, Markus, Matsumot, and Norasakkunkit (1997) analyzed 23 Asian studies and found a robust self-efficacy effect. That is, people in Asian cultures tended to attribute their own success externally and attribute their failures to a lack of ability and talent. They also found that Japanese employees accept failure better than they accept success. The authors suggested that the self is made meaningful in individualist cultures by reference to a set of attributes that are internal and bounded to the self, whereas in collectivist cultures, a self-critical view is held, which serves as a symbolic act of affirming one’s belongingness to the group. Menon, Morris, Chiu, and Hong (1999) presented similar findings, also concluding that the bias toward attributing success to personal attributes in Western cultures is robust, while the effect is markedly reduced in Asian collectivist societies. Building on attribution theory’s position that consistent behavior is attributed to stable causes, they suggested that collectivists attribute behavior to stable characteristics within the situation, whereas individualists attribute behavior to stable characteristics within the person. Morris and Peng (1994) made a similar point, concluding that Americans are person centered and Chinese are situation centered. These authors suggested that Americans view people as individual units: They can leave groups at any time and are socialized to behave according to personal preferences. Conversely, the Chinese view people as group members who cannot leave groups at will and must behave according to group norms, role constraints, and situational scripts. Consistent with this idea, these authors found that individualists are likely to reward short-term individual behaviors, and collectivists are likely to reward long-term group behaviors.

Given that most performance appraisal formats try to evaluate performance at the individual level, asking about specific behaviors, our methods of performance rating may not be appropriate in collectivist cultures. Attributional differences between individualist and collectivist cultures are deeply rooted in their respective dominant social representations (i.e., Judeo-Christian belief in individual soul and free will vs. Confucian primacy of social relationships and virtue of role-appropriate behavior) (Morris & Peng, 1994). For this reason, cultural differences in performance rating should not be expected to disappear through brief, local interventions such as rater training.

Research evidence on attribution styles suggests that it would be naive to expect high levels of agreement across individualist and collectivist raters. This theoretical argument is becoming increasingly more important as organizations move into international markets. As organizations become global, researchers must ask and answer questions designed to understand precisely how performance ratings differ across cultures.

**LEGAL ISSUES IN PERFORMANCE RATINGS**

Because performance ratings are used as a subjective basis for making personnel decisions, they are fair game in the legal arena. The performance appraisal process is often chal-
lenged in the courts (Malos, 1998). To gain an appreciation of the role of the legal system in performance ratings it is important to understand how the courts tend to decide in performance appraisal cases, as well as the laws regarding performance appraisal.

The most influential law relevant to performance appraisal is Title VII of the Civil Rights Act of 1964. This act made it unlawful to discriminate on the basis of race, color, religion, sex, or national origin when making administrative decisions (Gutman, 1993). Further extending the rules set by Title VII was the 1967 Age Discrimination Act, which prohibits discrimination against those over 40 years of age. Finally, in 1992 the Americans with Disabilities Act was passed, which prohibits discrimination on the basis of a person's disabilities.

Although these laws establish protected classes against whom personnel decisions must not discriminate, the laws do not address the definition of discrimination. In 1978, to help define when unfair discrimination occurs, the Equal Employment Opportunity Commission adopted the Uniform Guidelines on Employee Selection Procedures. Murphy and Cleveland (1995, p. 11) cited this as the "single greatest influence on the development and use of personnel assessment in the United States." The courts have turned to these guidelines when interpreting the laws discussed previously. The Uniform Guidelines specify criteria to be used in determining whether any subgroup has been discriminated against (Ledvinka & Scarpello, 1992).

In addition to the Guidelines, there have also been landmark court decisions relevant to the legal interpretation of performance ratings. Precedents set by these decisions must be considered when attempting to predict how judges will interpret ratings cases. For example, in Watson v. Fort Worth Bank and Trust (1988), the Supreme Court ruled that if subjective ratings are used for administrative decisions, they must be job related. Further, it was ruled that ratings are subject to the same statistical tests to determine discrimination as objective tests. In Abermarle Paper Company v. Moody (1975), the Supreme Court ruled that accuracy and validity of a performance appraisal process could be assessed in a validation study. Further, in Brito v. Zia (1973), the courts ruled that performance ratings are tests and must conform to the Uniform Guidelines (Cascio & Bernardin, 1981). Bersoff (1988) reiterated the suggestion that subjective ratings be validated in the same manner as objective tests.

In an alternative approach to the test metaphor, Folger, Konovsky, and Cropanzano (1992) presented performance appraisal using a due process metaphor. These authors suggested that the test metaphor is deficient because it assumes that work situations allow for reliable and valid measurement, which is probably not the case in our rapidly changing work environments. Also, as mentioned earlier, treating performance ratings as a test ignores the cognitive limitations and/ or goals adopted by the raters as a source of systematic rating variance. The due process metaphor presented by Folger et al. (1992) suggests that the legality of personnel decisions should be based on how the decisions are implemented (i.e., Did all parties involved have plenty of notice before the decision, voice in the decision, and confidence that the decision was based on evidence?).

Feild and Holley (1982) analyzed influential factors in 66 judicial decisions regarding performance appraisal between 1965 and 1980. These authors identified five important factors: (1) the use of job analysis, (2) the use of behavioral-oriented appraisal (as opposed to trait-based ratings of performance), (3) whether or not the raters were provided with adequate instruction, (4) whether or not the ratings would be reviewed by the ratee, and (5) the type of organization. In sum, these researchers concluded that nonindustrial organizations using behavioral rating scales based on job analyses are the most defensible, especially when raters are given the proper instruction and ratings are shown to the targets.

In a more recent review by Werner and Bolino (1997) both the test metaphor and the due process metaphor were examined regarding court decisions. Similar to the Feild and Holley (1982) study, these authors found that judges favorably considered job analysis, written rater instructions, and ratee review of ratings. Contradictory to Feild and Holley (1982), Werner and Bolino (1997) found that the type of organization and type of appraisal did not make a difference in court decisions. Further, appraisal systems incorporating multiple raters with adequate levels of agreement were preferable to single rater assessments. These researchers also found that the issues described regarding the due process metaphor are important to judges when evaluating appraisal cases. Finally, the authors found that validation concerns in subjective ratings were virtually ignored, which argues against the importance of the test metaphor. In support of this notion, Beck-Dudley and McEvoy (1991) found that the courts have a general hesitation to consider performance appraisal validity evidence.

Beck-Dudley and McEvoy (1991) also concluded that the courts do not clearly express to performance-rating consumers what the essential characteristics of appraisal systems are that make them legally defensible. The best available advice is that of Feild and Holley (1982) and Werner and Bolino (1997). For further insights into how to make an appraisal system legally defensible, the interested reader can refer to Ashe and McRae (1985), Barrett and Kernan (1987), Feild and Holley (1982), Malos (1998), Martin and Bartol (1991), Veglahn (1993), and Werner and Bolino (1997).
FUTURE DEVELOPMENTS

Both empirical research and theoretical development related
to performance ratings continue. We have selected for brief
discussion a few topics that we believe likely to have an
important impact in the near future.

Rater Goals

Murphy and Cleveland (1991, 1995) suggested that it may
be advantageous to think of the performance rating process
as a goal-directed communication process. These authors
noted that when evaluating performance, raters attempt to
use their ratings to communicate information consistent with
their personal interests. This idea is important because it im-
plies that raters are not always motivated to provide accurate
ratings.

Research on rater goals is in its infancy, and there are still
many unanswered questions about what types of goals raters
adopt, as well as the effects that those goals exert on ratings.
Some progress has been made, however. In a recent study
investigating the goals of raters, goals were found to influ-
ence performance ratings. Skattebo, Newman, Kinney, and
Cleveland (2002) found support for four goals adopted by
university students when rating instructors. Raters adopted
the goal of identifying strengths in their instructor’s perfor-
mance, identifying weaknesses in the instructor’s performance,
motivating the instructor, or rating fairly. These goals were
related to the ratings assigned by students. Specifically, to the
extent that a student endorsed “identify strength” goals, he
or she was more likely to assign high performance ratings for
the instructor. Also, to the extent that students endorsed “iden-
tify weakness” goals, they were likely to assign low ratings.
The results of this study suggest that the adoption of different
goals by the raters does relate to the level of the performance
rating.

The key point in considering the goals adopted by raters
is that if raters’ goals are not to rate accurately, raters will
not rate accurately regardless of the rater training, the rating
scale format, the cognitive judgment, and so on. As Bjørke,
Cleveland, Morrison, and Wilson (1987) pointed out, the
choice of exactly what message the rater communicates with
performance ratings depends on the goals being pursued by
such rater. If the impact of rater goals on performance ratings
is not considered, consumers of performance appraisal informa-
tion will not gain an accurate understanding of the messages
communicated by the performance rating process. Although
research has yet to answer a lot of the questions surrounding
rater goals, this area does promise to be a hot topic in the
future.

Rater Training and Cognitive Processes

As stated earlier in this chapter, in the 1980s and 1990s much
attention was focused on researching the cognitive processes
of raters. Through this research, an understanding of how
raters observe, encode, and retrieve performance information
has been gained. A criticism of this research is that the ad-
vances in understanding of rater cognitive processes have not
led to improvements in the practice of performance appraisal
(Murphy & Cleveland, 1995).

One area where this research could inform practice is the
area of rater training. As discussed earlier, several recent
advances have occurred in the methods used in training raters.
A number of these advances involve training raters on how
to process information. For example, FORT and BOT teach
raters what behaviors are important to encode, retrieve, and
recall. However, at this time not enough research has accum-
ulated about the effectiveness of these “cognitive training”
approaches.

FORT has received favorable evaluations (see Arvey &
Murphy, 1998; Bernardin et al., 2000; Woehr & Huffcutt,
1994), although Bernardin et al. (2000) noted that most
FORT research has been within laboratory settings. Given the
distinction between performance judgment and performance
appraisal suggested by Murphy and Cleveland (1995) and
explained at the beginning of this chapter, the restriction to
laboratory settings is an important consideration. Because
strict controls can be placed on ratings in the lab, the ratings
can exist in an environment less affected by contextual vari-
able. Therefore, ratings made in these settings are a closer
reflection of the rater’s judgments than could be expected in
a context-rich organizational setting.

Given these concerns, two areas for future research can be
identified. First, more information is needed to determine the
effectiveness of these “cognitively based” training methods
in applied settings. Second, if research in this area does show
that cognitively-based training methods are not as effective
outside the laboratory, a shift in focus of rater training may
be in order. Specifically, training could focus on preparing
raters to filter away context and make ratings that are con-
sistent with their judgments. Such an intervention would likely
address rater goals and organizational goals with the purpose
of minimizing the discrepancy between the two.

Rating Format Research

Recent research has called to question Landy and Farr’s
(1980) moratorium on rating format research on two grounds:
(1) the use of ratings for feedback purposes requires that rat-
ers be allowed to communicate global performance infor-

mation (consistent with memory categorization principles), whereas ratees desire feedback that is specific and behavioral (identifying relative personal strengths and weaknesses) (Jelley & Goffin, 2001), and (2) adaptive formats that are customized to fit the information-processing tendencies of individual raters (CARS) show greater accuracy than BARS or GRS (Borman et al., 2001). Although the contentions introduced by these studies mark the implicit promise of future research on the topic of rating formats, respective effect sizes for differential accuracy of various formats were small (typically less than .10). If future research on rating formats is to ever become justifiable, then it is essential for innovative new formats to be introduced through conscientious application of theories of rater cognition. Whether such custom alignment of rating formats to each rater’s cognitive processes will bear fruit remains to be seen.

**Toward a Multilevel Conceptualization of Performance**

Ultimately, human resource professionals and managers who use job performance ratings are interested in improving firm-level performance (DeNisi, 2000). Within the paradigm of performance appraisal research, it has often been assumed that individual-level improvements in performance will cumulate to produce firm-level improvement. This is exemplified by utility analysis (Schmidt, Hunter, McKenzie, & Muldrow, 1979). Utility estimates are problematic, however, because they ignore important processes by which individual performance aggregates (e.g., teamwork, information sharing between workgroups) and are thus subject to an *atomistic fallacy* of ignoring the lack of generality from an individual unit of analysis to a group or organizational unit of analysis (House, Rousseau, & Thomas-Hunt, 1995; Kozlowski & Klein, 2000).

On many occasions, strong individual-level performance does not enhance group-level performance, such as when a team is engaged in a sequentially interdependent task, for which group performance is determined by the “weakest link” rather than by the average performance of team members (Tesluk, Mathieu, Zaccaro, & Marks, 1999). Furthermore, performance rating is an inherently multilevel phenomenon when ratings are used to make between-persons comparisons for personnel decisions. When making such decisions as, “Which employee should be promoted?”, the basis for each individual decision must incorporate information about the performance of all group members (DeNisi, 2000). Our reason for pointing out the multilevel nature of the performance construct is to make clear that the ultimate imperative of *attaining organization-level objectives* should be recognized as the driving force behind every step in the design of the performance rating system.

**REFERENCES**


