Performance management is a “continuous process of identifying, measuring, and developing the performance of individuals and teams and aligning performance with the strategic goals of the organization” (Aguinis, 2019, p. 4). It is not a one-time event that takes place during the annual performance-review period. Rather, performance is assessed at regular intervals, and feedback is provided so that performance is improved on an ongoing basis. Performance appraisal is the systematic description of job-relevant strengths and weaknesses within and between employees or groups. It is a critical component of all performance management systems. Researchers and practitioners have been fascinated by how to measure and improve performance for decades; yet their overall inability to resolve definitively the knotty technical and interpersonal problems of performance appraisal and management has led one reviewer to term it the “Achilles heel” of human resource management (Heneman, 1975). This statement still applies today (DeNisi & Murphy, 2017). Supervisors and subordinates alike are intensely aware of the political and practical implications of the ratings and, in many cases, are acutely ill at ease during performance appraisal interviews. Despite these shortcomings, surveys of managers from both large and small organizations consistently show that they are unwilling to abandon performance management. For example, a survey of performance management systems and practices in 278 organizations across 15 countries found that about 90% use a company-sanctioned performance management system (Cascio, 2011).

Many treatments of performance management scarcely contain a hint of the emotional overtones, the human problems, so intimately bound up with it (Aguinis, 2019). Traditionally, researchers have placed primary emphasis on technical issues—for example, the advantages and disadvantages of various rating systems, sources of error, and problems of unreliability in performance observation and measurement (Aguinis & Pierce, 2008). To be sure, these are vitally important concerns. No less important, however, are the human issues involved, for performance management is not merely a technique—it is a process, a dialogue involving both people and data, and this process also includes social, motivational, and interpersonal aspects (Fletcher, 2001). In addition, performance management needs to be placed within the broader context of the organization’s vision, mission, and strategic priorities. A performance management system will not be successful if it is not linked explicitly to broader work unit and organizational goals.

In this chapter, we focus on both the measurement and the social/motivational aspects of performance management. As HR specialists, our task is to make the formal process as meaningful and workable as present research and knowledge will allow.

Purposes Served

Performance management systems that are designed and implemented well can serve several important purposes:

Performance management systems serve a strategic purpose because they help link employee activities with the organization's mission and goals. Well-designed performance management systems identify the behaviors and results needed to carry out the organization's strategic priorities and maximize the extent to which employees exhibit the desired behaviors and produce the intended results.

Performance management systems serve an important communication purpose because they allow employees to know how they are doing and what the organizational expectations are regarding their performance. They convey the aspects of work the supervisor and other organization stakeholders believe are important.

Performance management systems can serve as bases for employment decisions — decisions to promote outstanding performers; to terminate marginal or low performers; to train, transfer, or discipline others; and to award merit increases (or no increases). In short, information gathered by the performance management system can serve as predictors and, consequently, as key inputs for administering a formal organizational reward and punishment system (Cummings, 1973), including promotional decisions.

Data regarding employee performance can serve as criteria in HR research (e.g., in test validation).

Performance management systems also serve a developmental purpose because they can help establish objectives for training programs based on concrete feedback. To improve performance in the future, an employee needs to know what his or her weaknesses were in the past and how to correct them in the future. Pointing out strengths and weaknesses is a coaching function for the supervisor; receiving meaningful feedback and acting on it constitute a motivational experience for the subordinate. Thus, performance management systems can serve as vehicles for personal development.

Performance management systems can facilitate organizational diagnosis, maintenance, and development. Proper specification of performance levels, in addition to suggesting training needs across units and indicating necessary skills to be considered when hiring, is important for HR planning and HR evaluation. It also establishes the more general organizational requirement of ability to discriminate effective from ineffective performers. Appraising employee performance, therefore, represents the beginning of a process rather than an end product (Jacobs, Kafry, & Zedeck, 1980).

Finally, performance management systems allow organizations to keep proper records to document HR decisions and legal requirements.

Realities and Challenges of Performance Management Systems

Independent of any organizational context, the implementation of performance management systems at work confronts organizations with five realities (Ghorpade & Chen, 1995):

This activity is inevitable in all organizations, large and small, public and private, and domestic and multinational. Organizations need to know if individuals are performing competently, and, in the current legal climate, appraisals are essential features of an organization's defense against challenges to adverse employment actions, such as terminations or layoffs.

Appraisal is fraught with consequences for individuals (rewards and punishments) and organizations (the need to provide appropriate rewards and punishments based on performance).

As job complexity increases, it becomes progressively more difficult, even for well-meaning appraisers, to assign accurate, merit-based performance ratings.

When evaluating coworkers, there is an ever-present danger of the parties being influenced by the political consequences of their actions—rewarding allies and punishing enemies or competitors (Longenecker & Gioia, 1994).

The implementation of performance management systems takes time and effort, and participants (those who rate performance and those whose performance is rated) must be convinced the system is useful and fair. Otherwise, the system may carry numerous negative consequences. For example, high-performing employees may quit, time and money may be wasted, and adverse legal consequences may result.

Overall, these five realities involve several political and interpersonal challenges. Political challenges stem from deliberate attempts by raters to enhance or to protect their self-interests when conflicting courses of action are possible. Political considerations are facts of organizational life (Westphal & Clement, 2008). Appraisals take place in an organizational environment that is anything but completely rational, straightforward, or dispassionate. It appears that achieving accuracy in appraisal is less important to managers than motivating and rewarding their subordinates. Many managers will not allow excessively accurate ratings to cause problems for themselves, and they attempt to use the appraisal process to their own advantage. Interpersonal challenges arise from the actual face-to-face encounter between subordinate and superior. Because of a lack of communication, employees may think they are being judged according to one set of standards when their superiors actually use different ones. Furthermore, supervisors often delay or resist making face-to-face appraisals. Rather than confronting substandard performers with low ratings, negative feedback, and below-average salary increases, supervisors often find it easier to "damn with faint praise" by giving average or above-average ratings to inferior performers (Benedict & Levine, 1988). Finally, some managers complain that formal performance appraisal interviews tend to interfere with the more constructive coaching relationship that should exist between superior and subordinate. They claim that appraisal interviews emphasize the superior position of the supervisor by placing him or her in the role of judge, which conflicts with the supervisor's equally important roles of teacher and coach (Meyer, 1991).

This, then, is the performance management dilemma: It is widely accepted as a potentially useful tool, but political and interpersonal barriers often thwart its successful implementation. There is currently an intense debate in both research and practitioner circles on how to solve this dilemma. In recent years, some large organizations including Accenture, Deloitte, Microsoft, Gap, Inc., and Eli Lilly chose to abandon or substantially curtail their use of performance appraisal (Adler et al., 2016), but most of them later realized that appraisals are critical given the purposes listed earlier (Hunt, 2016).

Much of the research on appraisals has focused on measurement issues. This is important, but HR professionals may contribute more by improving the attitudinal and interpersonal components of performance appraisal systems, as well as their technical aspects. Let's begin by considering the fundamental requirements for a best-in-class performance management system.

Fundamental Requirements of Successful Performance Management Systems

For any performance management system to be used successfully, it should have the following nine characteristics (Aguinis, 2019; Aguinis, Joo, & Gottfredson, 2011):

Congruence with strategy: The system should measure and encourage behaviors that will help achieve organizational goals.

Thoroughness: All employees should be evaluated, all key job-related responsibilities should be measured, and evaluations should cover performance for the entire time period included in any specific review.

Practicality: The system should be available, plausible, acceptable, and easy to use, and its benefits should outweigh its costs.

Meaningfulness: Performance measurement should include only matters under the employee's control, appraisals should occur at regular intervals, the system should provide for continuing skill development of raters and ratees, results should be used for important HR decisions, and implementation of the system should be seen as an important part of everyone's job.

Specificity: The system should provide specific guidance to both raters and ratees about what is expected of them and also how they can meet these expectations.

Discriminability: The system should allow for clear differentiation between effective and ineffective performance and performers.

Reliability and validity: Performance scores should be consistent over time and across raters observing the same behaviors (see Chapter 6) and should not be deficient or contaminated (see Chapter 4).

Inclusiveness: Successful systems allow for the active participation of raters and ratees, including in the design of the system (Kleingeld, Van Tuijl, & Algera, 2004). This includes allowing ratees to provide their own performance evaluations and to assume an active role during the appraisal interview, and allowing both raters and ratees an opportunity to provide input in the design of the system.

Fairness and acceptability: Participants should view the process and outcomes of the system as being just and equitable.

Several studies have investigated these characteristics, which dictate the success of performance management systems (Cascio, 1982). For example, regarding meaningfulness, a study including 176 Australian government workers indicated that the system's meaningfulness (i.e., perceived consequences of implementing the system) was an important predictor of the decision to adopt or reject a system (Langan-Fox, Waycott, Morizzi, & McDonald, 1998). Regarding inclusiveness, a meta-analysis of 27 studies, including 32 individual samples, found that the overall correlation between employee participation and employee reactions to the system (corrected for unreliability) was .61 (Cawley, Keeping, & Levy, 1998). Specifically, the benefits of designing a system in which ratees are given a "voice" included increased satisfaction with the system, increased perceived utility of the system, increased motivation to improve performance, and increased perceived fairness of the system (Cawley et al., 1998).

Taken together, the nine key characteristics indicate that performance appraisal should be embedded in the broader performance management system and that a lack of understanding of the context surrounding the appraisal is likely to result in a failed system. With that in mind, let's consider the benefits of state-of-the-science performance management systems.

Benefits of State-of-the-Science Performance Management Systems

When performance management systems are implemented following the requirements described in the previous section, they can be a clear source of competitive advantage (Aguinis, Joo, & Gottfredson, 2011). Specifically, such state-of-the-science systems benefit employees, managers, and organizations. For example, as shown in Table 5.1, employees understand what is expected of them and learn about their own strengths and weaknesses, which is useful information for their own personal development. Similarly, managers obtain insights regarding their subordinates and are able to obtain more precise and differentiating information that is necessary for making administrative decisions (e.g., promotions, compensation decisions), as well as for creating personal development plans. Finally, organizations are able to implement policies that are fair, standardized, and acceptable. Overall, the way to solve the dilemma mentioned earlier is not to get rid of performance appraisal and management, but to implement systems following best-practice recommendations based on the available empirical evidence.

Who Shall Rate?

In view of the purposes served by performance management, who does the rating is important. In addition to being cooperative and trained in the techniques of rating, raters must have direct experience with, or firsthand knowledge of, the individual to be rated. In many jobs, individuals with varying perspectives have such firsthand knowledge. Following are descriptions of five of these perspectives that will help answer the question of who shall rate performance.

Immediate Supervisor

The supervisor is probably the person best able to evaluate each subordinate's performance in light of the organization's overall objectives. Since the supervisor is probably also responsible for reward (and punishment) decisions such as pay, promotion, and discipline, he or she must be able to tie effective (ineffective) performance to the employment actions taken. Inability to form such linkages between performance and punishment or reward is one of the most serious deficiencies of any performance management system.

However, in jobs such as teaching, law enforcement, or sales and in self-managed work teams, the supervisor may only rarely observe his or her subordinate's performance directly. In addition, performance ratings provided by the supervisor may reflect not only whether an employee is helping advance organizational objectives but also whether the employee is contributing to goals valued by the supervisor, which may or may not be congruent with organizational goals (Hogan & Shelton, 1998). Moreover, if a supervisor has recently received a positive evaluation regarding his or her own performance, he or she is also likely to provide a positive evaluation regarding his or her subordinates

(Latham, Budworth, Yanar, & Whyte, 2008). Fortunately, several other perspectives can be used to provide a fuller picture of the individual's total performance.

Peers

Peer assessment refers to three of the more basic methods used by members of a well-defined group in judging each other's job performance. These include peer nominations, most useful for identifying persons with extreme high or low levels of performance; peer rating, most useful for providing feedback; and peer ranking, best at discriminating various levels of performance from highest to lowest on each dimension.

Reviews of peer assessment methods reached favorable conclusions regarding the reliability, validity, and freedom from biases of this source of performance information (e.g., Kane & Lawler, 1978). However, some problems still remain. First, two characteristics of peer assessments appear to be related significantly and independently to user acceptance (McEvoy & Buller, 1987). Perceived friendship bias is related negatively to user acceptance, and use for developmental purposes is related positively to user acceptance. How do people react upon learning that they have been rated poorly (favorably) by their peers? Research in a controlled setting indicates that such knowledge has predictable effects on group behavior. Negative peer-rating feedback produces significantly lower perceived performance of the group, plus lower cohesiveness, satisfaction, and peer ratings on a subsequent task. Positive peer-rating feedback produces nonsignificantly higher values for these variables on a subsequent task (DeNisi, Randolph, & Blencoe, 1983). One possible solution that might simultaneously increase feedback value and decrease the perception of friendship bias is to specify clearly (e.g., using critical incidents) the performance criteria on which peer assessments are based. Results of the peer assessment may then be used in joint employee–supervisor reviews of each employee's progress, prior to later administrative decisions concerning the employee.

A second problem with peer assessments is that they seem to include more common method variance than assessments provided by other sources. Method variance is the variance observed in a performance measure that is not relevant to the behaviors assessed, but instead is due to the method of measurement used (Brannick, Chan, Conway, Lance, & Spector, 2010; Conway, 2002). For example, Conway (1998) reanalyzed supervisor, peer, and self-ratings for three performance dimensions (i.e., altruism-local, conscientiousness, and altruism-distant) and found that the proportion of method variance for peers was .38, whereas the proportion of method variance for self-ratings was .22. This finding suggests that relationships among various performance dimensions, as rated by peers, can be inflated substantially due to common method variance (Conway, 1998).

Several data-analysis methods are available to estimate the amount of method variance present in a peer-assessment measure (Conway, 1998; Williams, Hartman, & Cavazotte, 2010). At the very least, the assessment of common method variance can provide HR researchers and practitioners with information regarding the extent of the problem. In addition, Podsakoff, MacKenzie, Lee, and Podsakoff (2003) proposed two types of remedies to address this problem:

Procedural remedies: These include obtaining measures of the predictor and criterion variables from different sources; separating the measurement of the predictor and criterion variables (i.e., temporal, psychological, or methodological separation); protecting respondent anonymity, thereby reducing socially desirable responding; counterbalancing the question order; and improving scale items.

Statistical remedies: These include utilizing Harman's single-factor test (i.e., to determine whether all items load into one common underlying factor, as opposed to the various factors hypothesized); computing partial correlations (e.g., partialling out social desirability, general affectivity, or a general factor score); controlling for the effects of a directly measured latent methods factor; controlling for the effects of a single, unmeasured, latent method factor; implementing the correlated uniqueness model (i.e., where a researcher identifies the sources of method variance so that the appropriate pattern of measurement-error corrections can be estimated); and utilizing the direct-product model (i.e., which models trait-by-method interactions).

The overall recommendation is to follow all the procedural remedies listed here, but the statistical remedies to be implemented depend on the specific characteristics of the situation faced (Podsakoff et al., 2003).

Given our discussion thus far, peer assessments are probably best considered as only one element in a system that includes input from all sources that have unique information or perspectives to offer. Thus, the behaviors and outcomes to be assessed should be considered in the context of the groups and situations in which peer assessments are to be applied. It is impossible to specify, for all situations, the kinds of characteristics that peers are able to rate best.

Subordinates

Subordinates offer a somewhat different perspective on a manager's performance. They know directly the extent to which a manager does or does not delegate, the extent to which he or she plans and organizes, the type of leadership style(s) he or she is most comfortable with, and how well he or she communicates. This is why subordinate ratings often provide information that accounts for variance in performance measures over and above other sources (Conway, Lombardo, & Sanders, 2001). This approach is used regularly by universities (students evaluate faculty) and sometimes by large corporations, where a manager may have many subordinates. In small organizations, however, considerable trust and openness are necessary before subordinate appraisals can pay off.

They can pay off, though. For example, a study in a public institution with about 2,500 employees that performs research, development, tests, and evaluation in South Korea provided evidence of the benefits of upward appraisals—particularly long-term benefits (Jhun, Bae, & Rhee, 2012). Functional managers received upward feedback once a year during a period of seven years. For purposes of the analysis, they were divided into low, medium, and high performers. Results showed that those in the low-performing group benefited the most. Moreover, when upward feedback was used for administrative rather than developmental purposes, the impact on performance improvement was even larger.

Subordinate ratings have been found to be valid predictors of subsequent supervisory ratings over two-, four-, and seven-year periods (McEvoy & Beatty, 1989). One reason for this may have been that multiple ratings on each dimension were made for each manager, and the ratings were averaged to obtain the measure for the subordinate perspective. Averaging has several advantages. First, averaged ratings are more reliable than single ratings. Second, averaging helps to ensure the anonymity of the subordinate raters. Anonymity is important; subordinates may perceive the process to be threatening, since the supervisor can exert administrative controls (salary increases, promotions, etc.). In fact, when the identity of subordinates is disclosed, inflated ratings of managers' performance tend to result (Antonioni, 1994).

Any organization contemplating use of subordinate ratings should pay careful attention to the intended purpose of the ratings. Evidence indicates that ratings used for salary administration or promotion purposes may be more lenient than those used for guided self-development (Zedeck & Cascio, 1982). In general, subordinate ratings are of significantly better quality when used for developmental purposes rather than administrative purposes (Greguras, Robie, Schleicher, & Goff, 2003).

Self

It seems reasonable to have each individual judge his or her own job performance. On the positive side, we can see that the opportunity to participate in performance appraisal, especially if it is combined with goal setting, should improve the individual's motivation and reduce his or her defensiveness during an appraisal interview. Research to be described later in this chapter clearly supports this view. On the negative side, comparisons with appraisals by supervisors, peers, and subordinates suggest that self-appraisals tend to show more leniency, less variability, more bias, and less agreement with the judgments of others (Atkins & Wood, 2002; Harris & Schaubroeck, 1988). This seems to be the norm in Western cultures. In Taiwan, however, modesty bias (self-ratings lower than those of supervisors) has been found (Farh, Dobbins, & Cheng, 1991), although this may not be the norm in all Eastern cultures (Barron & Sackett, 2008).

To some extent, idiosyncratic aspects of self-ratings may stem from the tendency of raters to base their scores on different aspects of job performance or to weight facets of job performance differently. Self- and supervisor ratings agree much more closely when both parties have a thorough knowledge of the appraisal system or process (Williams & Levy, 1992). In addition, self-ratings are less lenient when done for self-development purposes rather than for administrative purposes (Meyer, 1991). In addition, self-ratings of contextual performance are more lenient than peer ratings when individuals are high on self-monitoring (i.e., tending to control self-presentational behaviors) and social desirability (i.e., tending to attempt to make oneself look good) (Mersman & Donaldson, 2000). The situation is far from hopeless, however. To improve the validity of self-appraisals, consider four research-based suggestions (Campbell & Lee, 1988; Fox & Dinur, 1988; Mabe & West, 1982):

Instead of asking individuals to rate themselves on an absolute scale (e.g., a scale ranging from "poor" to "average"), provide a relative scale that allows them to compare their performance with that of others

(e.g., “below average,” “average,” “above average”). In addition, providing comparative information on the relative performance of coworkers promotes closer agreement between self-appraisal and supervisor rating (Farh & Dobbins, 1989).

Provide multiple opportunities for self-appraisal, for the skill being evaluated may well be one that improves with practice.

Provide reassurance of confidentiality—that is, that self-appraisals will not be “publicized.”

Focus on the future—specifically on predicting future behavior.

Until the problems associated with self-appraisals can be resolved, however, they seem more appropriate for counseling and development than for employment decisions.

Clients Served

Another group that may offer a different perspective on individual performance in some situations is that of clients served. In jobs that require a high degree of interaction with the public or with particular individuals (e.g., purchasing managers, suppliers, and sales representatives), appraisal sometimes can be done by the consumers of the organization’s services. Although the clients served cannot be expected to identify completely with the organization’s objectives, they can, nevertheless, provide useful information. Such information may affect employment decisions (promotion, transfer, need for training), but it also can be used in HR research (e.g., as a criterion in validation studies or in the measurement of training outcomes on the job) or as a basis for self-development activities.

Appraising Performance: Individual Versus Group Tasks

So far, we have assumed that ratings are assigned on an individual basis. That is, each source—be it the supervisor, peer, subordinate, self, or client—makes the performance judgment individually and independently from other individuals. However, in practice, appraising performance is not strictly an individual task. A survey of 135 raters from six organizations indicated that 98.5% of raters reported using at least one secondhand (i.e., indirect) source of performance information (Raymark, Balzer, & De La Torre, 1999). In other words, supervisors often use information from outside sources in making performance judgments. Moreover, supervisors may change their own ratings in the presence of indirect information. For example, a study including participants with at least two years of supervisory experience revealed that supervisors are likely to change their ratings when the ratee’s peers provide information perceived as useful (Makiney & Levy, 1998). A follow-up study that included students from a Canadian university revealed that indirect information is perceived to be most useful when it is in agreement with the rater’s direct observation of the employee’s performance (Uggerslev & Sulsky, 2002). For example, when a supervisor’s judgment about a ratee’s performance is positive, positive indirect observation produced higher ratings than negative indirect information. In addition, it seems that the presence of indirect information is more likely to change ratings from positive to negative than from negative to positive (Uggerslev & Sulsky, 2002). In sum, although direct observation is the main influence on ratings, the presence of indirect information is likely to affect ratings.

If the process of assigning performance ratings is not entirely an individual task, might it pay off to formalize performance appraisals as a group task? One study found that groups are more effective than individuals at remembering specific behaviors over time, but that groups also demonstrate greater response bias (Martell & Borg, 1993). In a second related study, individuals observed a 14-minute military training videotape of five men attempting to build a bridge of rope and planks in an effort to get themselves and a box across a pool of water. Before observing the tape, study participants were given indirect information in the form of a positive or negative performance cue [i.e., "the group you will observe was judged to be in the top (bottom) quarter of all groups"]. Then ratings were provided individually or in the context of a four-person group (the group task required that the four group members reach consensus). Results showed that ratings provided individually were affected by the performance cue, but that ratings provided by the groups were not (Martell & Leavitt, 2002).

These results suggest that groups can be of help, but they are not a cure-all for the problems of rating accuracy. Groups can be a useful mechanism for improving the accuracy of performance appraisals under two conditions. First, the task needs to have a necessarily correct answer. For example, is the behavior present or not? Second, the magnitude of the performance cue should not be too large. If the performance facet in question is subjective (e.g., "what is the management potential for this employee?") and the magnitude of the performance cue is large, group ratings may amplify instead of attenuate individual biases (Martell & Leavitt, 2002).

In summary, there are several sources of appraisal information, and each provides a different perspective, a different piece of the puzzle. The various sources and their potential uses are shown in Table 5.2. Several studies indicate that data from multiple sources (e.g., self, supervisors, peers, subordinates) are desirable because they provide a complete picture of the individual's effect on others (Borman, White, & Dorsey, 1995; Murphy & Cleveland, 1995; Wohlers & London, 1989).

Table 5.2 Sources and Uses of Appraisal Data

Use

Source

Supervisor

Peers

Subordinates

Self

Clients Served

Employment decisions

x

–

x

–

x

Self-development

x

x

x

x

x

HR research

x

x

_

_

x

Putting It All Together: 360-Degree Systems

As is obvious by now, the different sources of performance information are not mutually exclusive. So-called 360-degree feedback systems broaden the base of appraisals by including input from self, peers, subordinates, and (in some cases) clients. Moreover, there are several advantages to using these systems compared to a single source of performance information (Campion, Campion, & Campion, 2015). First, 360-degree feedback systems result in improved reliability of performance information because it originates from multiple sources and not just one source. Second, they consider a broader range of performance information, which is particularly useful in terms of minimizing criterion deficiency (as discussed in Chapter 4). Third, they usually include information not only on task performance but also on contextual performance and counterproductive work behaviors, which are all important given the multidimensional nature of performance. Finally, because multiple sources and individuals are involved, 360-degree systems have great potential to decrease biases—particularly compared to systems involving a single source of information.

For such systems to be effective, however, it is important to consider the following issues (Bracken & Rose, 2011):

Relevant content: The definition of success, no matter which is the source, needs to be clear and aligned with strategic organizational goals.

Data credibility: Each source needs to be perceived as capable and able to assess the performance dimensions assigned to it.

Accountability: Each participant in the system needs to be motivated to provide reliable and valid information—to the best of his or her ability.

Participation: Successful systems are typically implemented organizationwide rather than in specific units. This type of implementation will also facilitate acceptance.

Agreement and Equivalence of Ratings Across Sources

To assess the degree of interrater agreement within rating dimensions (convergent validity) and to assess the ability of raters to make distinctions in performance across dimensions (discriminant validity), a matrix listing dimensions as rows and raters as columns might be prepared (Lawler, 1967). As we noted earlier, however, multiple raters for the same individual may be drawn from different organizational levels, and they probably observe different facets of a ratee's job performance (Bozeman, 1997). This may explain, in part, why the overall correlation between subordinate and self-ratings (corrected for unreliability) is only .14, the correlation between subordinate and supervisor ratings (also corrected for unreliability) is .22 (Conway & Huffcutt, 1997), and the correlation between self and supervisory ratings is also only .22 (Heidemeier & Moser, 2009). Hence, having interrater agreement for ratings on all performance dimensions across organizational levels not only is an unduly severe expectation but also may be erroneous. Although we should not always expect agreement, we should expect that the construct underlying the measure used should be equivalent across raters. In other words, does the underlying trait measured across sources relate to observed rating scale scores in the same way across sources? In general, it does not make sense to assess the extent of interrater agreement without first establishing measurement equivalence (also called measurement invariance) because a lack of agreement may be due to a lack of measurement equivalence (Cheung, 1999). A lack of measurement equivalence means that the underlying characteristics being measured are not on the same psychological measurement scale, which implies that differences across sources are possibly artifactual, contaminated, or misleading (Maurer, Raju, & Collins, 1998).

Fortunately, there is evidence that measurement equivalence is present in many appraisal systems. Specifically, measurement equivalence was found in a measure of managers' team-building skills as assessed by peers and subordinates (Maurer, Raju, & Collins, 1998). Equivalence was also found in a measure including 48 behaviorally oriented items designed to measure 10 dimensions of managerial performance as assessed by self, peers, supervisors, and subordinates (Facteau & Craig, 2001) and in a meta-analysis including measures of overall job performance, productivity, effort, job knowledge, quality, and leadership as rated by supervisors and peers (Viswesvaran, Schmidt, & Ones, 2002). However, lack of equivalence was found for measures of interpersonal competence, administrative competence, and compliance and acceptance of authority as assessed by supervisors and peers (Viswesvaran et al., 2002). At this point, it is not clear what may account for differential measurement equivalence across studies and constructs, and this is a fruitful avenue for future research. One possibility is that behaviorally based ratings provided for developmental purposes are more likely to be equivalent than those reflecting broader behavioral dimensions (e.g., interpersonal competence) and collected for research purposes (Facteau & Craig, 2001). One conclusion is clear, however: Measurement equivalence needs to be established before ratings can be assumed to be directly comparable. Several methods exist for this purpose, including those based on confirmatory factor analysis (CFA) and item response theory (Barr & Raju, 2003; Cheung & Rensvold, 1999, 2002; Maurer, Raju, & Collins, 1998; Vandenberg, 2002).

Once measurement equivalence has been established, we can assess the extent of agreement across raters. For this purpose, raters may use a hybrid multitrait–multirater analysis (see Figure 5.1), in which raters make evaluations only on those dimensions that they are in good position to rate (Borman, 1974) and that reflect measurement equivalence. In the hybrid analysis, within-level interrater agreement is taken as an index of convergent validity. The hybrid matrix provides an improved conceptual fit for analyzing performance ratings, and the probability of obtaining convergent and discriminant validity is probably higher for this method than for the traditional multitrait–multirater analysis.

Figure 5.1 Example of a Hybrid Matrix Analysis of Performance Ratings

Note: Level I rates only traits 1–4. Level II rates only traits 5–8.

Another approach for examining performance ratings from more than one source is based on CFA (Williams et al., 2010). CFA allows researchers to specify each performance dimension as a latent factor and assess the extent to which these factors are correlated with each other. In addition, CFA allows for an examination of the relationship between each latent factor and its measures, as provided by each source (e.g., supervisor, peer, self). One advantage of using a CFA approach to examine ratings from multiple sources is that it allows for a better understanding of source-specific method variance (i.e., the dimension-rating variance specific to a particular source).

Judgmental Biases in Rating

In the traditional view, judgmental biases result from some systematic measurement error on the part of a rater. As such, they are easier to deal with than errors that are unsystematic or random. However, each type of bias has been defined and measured in different ways in the literature. This may lead to diametrically opposite conclusions, even in the same study (Saal, Downey, & Lahey, 1980). In the minds of many managers, however, these behaviors are not errors at all. For example, in an organization in which a team-based culture exists, can we really say that if peers place more emphasis on contextual than task performance in evaluating others, this is an error that should be minimized or even eliminated (cf. Lievens, Conway, & De Corte, 2008)? Rather, this apparent error is really capturing an important contextual variable in this particular type of organization. With these considerations in mind, let's consider some of the most commonly observed judgmental biases, along with ways of minimizing them.

Leniency and Severity

The use of ratings rests on the assumption that the human observer is capable of some degree of precision and some degree of objectivity (Guilford, 1954). His or her ratings are taken to mean something accurate about certain aspects of the person rated. "Objectivity" is the major hitch in these assumptions, and it is the one most often violated. Raters subscribe to their own sets of assumptions (that may or may not be valid), and most people have encountered raters who seemed either inordinately easy (lenient) or inordinately difficult (severe). Evidence also indicates that leniency is a

stable response tendency across raters (Kane, Bernardin, Villanova, & Peyrfitte, 1995). Moreover, some raters are more lenient than others, even in situations where there is little or no contact between raters and ratees after the performance evaluation (Dewberry, Davies-Muir, & Newell, 2013).

Senior managers recognize that leniency is not to be taken lightly. Fully 77% of sampled Fortune 100 companies reported that lenient appraisals threaten the validity of their appraisal systems (Bretz, Milkovich, & Read, 1990). An important cause for lenient ratings is the perceived purpose served by the performance management system in place. A meta-analysis that included 22 studies and a total sample size of more than 57,000 individuals concluded that when ratings are to be used for administrative purposes, scores are one third of a standard deviation larger than those obtained when the main purpose is research (e.g., validation study) or employee development (Jawahar & Williams, 1997). This difference is even larger when ratings are made in field settings (as opposed to lab settings), provided by practicing managers (as opposed to students), and provided for subordinates (as opposed to superiors). In other words, ratings tend to be more lenient when they have real consequences in actual work environments.

Leniency and severity biases can be controlled or eliminated in several ways: (a) by allocating ratings into a forced distribution, in which ratees are apportioned according to an underlying distribution (e.g., 20% of As, 70% of Bs, and 10% of Cs); (b) by requiring supervisors to rank order their subordinates; (c) by encouraging raters to provide feedback on a regular basis, thereby reducing rater and ratee discomfort with the process; and (d) by increasing raters' motivation to be accurate by holding them accountable for their ratings. For example, firms such as IBM, Pratt & Whitney, and Grumman implemented forced distributions because the extreme leniency in their ratings-based appraisal data hindered their ability to implement downsizing based on merit (Kane & Kane, 1993). Forced-distribution systems have their own disadvantages, however, as we describe later in this chapter.

Central Tendency

When political considerations predominate, raters may assign all their subordinates ratings that are neither too good nor too bad. They avoid using the high and low extremes of rating scales and tend to cluster all ratings about the center of all scales. "Everybody is average" is one way of expressing the central tendency bias. The unfortunate consequence, as with leniency or severity biases, is that most of the value of systematic performance appraisal is lost. The ratings fail to discriminate either within people over time or between people, and the ratings become virtually useless as managerial decision-making aids, as predictors, as criteria, or as a means of giving feedback.

Central tendency biases can be minimized by specifying clearly what the various anchors mean. In addition, raters must be convinced of the value and potential uses of merit ratings if they are to provide meaningful information.

Halo

Halo is perhaps the most actively researched bias in performance appraisal. A rater who is subject to the halo bias assigns ratings on the basis of a general impression of the ratee. An individual is rated either

high or low on specific factors because of the rater's general impression (good–poor) of the ratee's overall performance (Lance, LaPointe, & Stewart, 1994). According to this theory, the rater fails to distinguish among levels of performance on different performance dimensions. Ratings subject to the halo bias show spuriously high positive intercorrelations (Cooper, 1981).

Two critical reviews of research in this area (Balzer & Sulsky, 1992; Murphy, Jako, & Anhalt, 1993) led to the following conclusions: (a) Halo is not as common as believed; (b) the presence of halo does not necessarily detract from the quality of ratings (i.e., halo measures are not strongly interrelated, and they are not related to measures of rating validity or accuracy); (c) it is impossible to separate true from illusory halo in most field settings; and (d) although halo may be a poor measure of rating quality, it may or may not be an important measure of the rating process. So, contrary to assumptions that have guided halo research since the 1920s, it is often difficult to determine whether halo has occurred, why it has occurred (whether it is due to the rater or to contextual factors unrelated to the rater's judgment), or what to do about it. To address this problem, Solomonson and Lance (1997) designed a study in which true halo was manipulated as part of an experiment, and, in this way, they were able to examine the relationship between true halo and rater error halo. Results indicated that the effects of rater error halo were homogeneous across a number of distinct performance dimensions, although true halo varied widely. In other words, true halo and rater error halo are, in fact, independent. Therefore, the fact that performance dimensions are sometimes intercorrelated may not mean that there is rater bias but, rather, that there is a common, underlying general performance factor. Further research is needed to explore this potential generalized performance dimension.

As we noted earlier, judgmental biases may stem from a number of factors. One factor that has received considerable attention over the years has been the type of rating scale used. Each type attempts to reduce bias in some way. Although no single method is free of flaws, each has its own particular strengths and weaknesses. In the following section, we examine some of the most popular methods of evaluating individual job performance.

Types of Performance Measures

Objective Measures

Related to our discussion of performance as behaviors or results in Chapter 4, performance measures may be classified into two general types: objective and subjective. Objective performance measures include production data (dollar volume of sales, units produced, number of errors, amount of scrap) and employment data (accidents, turnover, absences, tardiness). Objective measures are usually, but not always, related to results. These variables directly define the goals of the organization and, therefore, sometimes are outside the employee's control. For example, dollar volume of sales is influenced by numerous factors beyond a particular salesperson's control—territory location, number of accounts in the territory, nature of the competition, distances between accounts, price and quality of the product, and so forth. This is why general cognitive ability scores predict ratings of sales performance quite well

(i.e., r = .40) but not objective sales performance (i.e., r = .04) (Vinchur, Schippmann, Switzer, & Roth, 1998).

Although objective measures of performance are intuitively attractive, they carry theoretical and practical limitations. But, because correlations between objective and subjective measures are far from being perfectly correlated (r = .39; Bommer, Johnson, Rich, Podsakoff, & Mackenzie, 1995), objective measures can offer useful information.

Subjective Measures

The disadvantages of objective measures have led researchers and managers to place major emphasis on subjective measures of job performance, which depend on human judgment. Hence, they are prone to the kinds of biases that we discuss in Chapter 6. To be useful, they must be based on a careful analysis of the behaviors viewed as necessary and important for effective job performance.

There is enormous variation in the types of subjective performance measures used by organizations. Some organizations use a long list of elaborate rating scales, others use only a few simple scales, and still others require managers to write a paragraph or two concerning the performance of each of their subordinates. In addition, subjective measures of performance may be relative (in which comparisons are made among a group of ratees) or absolute (in which a ratee is described without reference to others). In the next section, we briefly describe alternative formats.

Rating Systems: Relative and Absolute

We can classify rating systems into two types: relative and absolute. Within this taxonomy, the following methods may be distinguished:

Results of an experiment in which undergraduate students rated the videotaped performance of a lecturer suggest that no advantages are associated with the absolute methods (Wagner & Goffin, 1997). By contrast, relative ratings based on various rating dimensions (as opposed to a traditional global performance dimension) seem to be more accurate with respect to differential accuracy (i.e., accuracy in discriminating among ratees within each performance dimension) and stereotype accuracy (i.e., accuracy in discriminating among performance dimensions averaging across ratees). Given that the affective, social, and political factors influencing performance management systems were absent in this experiment conducted in a laboratory setting, view the results with caution. Moreover, a more recent study involving two separate samples found that absolute formats are perceived as fairer than relative formats (Roch, Sternburgh, & Caputo, 2007).

Because both relative and absolute methods are used pervasively in organizations, next we discuss each of these two types of rating systems in detail.

Relative Rating Systems (Employee Comparisons)

Employee comparison methods are easy to explain and are helpful in making employment decisions. They also provide useful criterion data in validation studies, for they effectively control leniency,

severity, and central tendency bias. Like other systems, however, they suffer from several weaknesses that should be recognized.

Employees usually are compared only in terms of a single overall suitability category. The rankings, therefore, lack behavioral specificity and may be subject to legal challenge. In addition, employee comparisons yield only ordinal data—data that give no indication of the relative distance between individuals. Moreover, it is often impossible to compare rankings across work groups, departments, or locations. The last two problems can be alleviated, however, by converting the ranks to normalized standard scores that form an approximately normal distribution. An additional problem is related to reliability. Specifically, when asked to rerank all individuals at a later date, the extreme high or low rankings probably will remain stable, but the rankings in the middle of the scale may shift around considerably.

Rank Ordering

Simple ranking requires only that a rater order all ratees from highest to lowest, from "best" employee to "worst" employee. Alternation ranking requires that the rater initially list all ratees on a sheet of paper. From this list, the rater first chooses the best ratee (#1), then the worst ratee (#n), then the second best (#2), then the second worst (#n−1), and so forth, alternating from the top to the bottom of the list until all ratees have been ranked.

Paired Comparisons

Both simple ranking and alternation ranking implicitly require a rater to compare each ratee with every other ratee, but systematic ratee-to-ratee comparison is not a built-in feature of these methods. For this, we need paired comparisons. The number of pairs of ratees to be compared may be calculated from the formula [n(n−1)]/2. Hence, if 10 individuals were being compared, [10(9)]/2 or 45 comparisons would be required. The rater's task is simply to choose the better of each pair, and each individual's rank is determined by counting the number of times he or she was rated superior.

Forced Distribution

In forced-distribution systems, raters must distribute a predetermined percentage of employees into categories based on their performance relative to other employees. This type of system results in a clear differentiation among groups of employees and became famous after legendary GE CEO Jack Welch implemented what he labeled the "vitality curve," in which supervisors identified the "top 20%," "vital 70%," and "bottom 10%" of performers within each unit. A recent literature review of the effects of forced-distribution systems concluded that they are particularly beneficial for jobs that are very autonomous (i.e., employees perform their duties without much interdependence) (Moon, Scullen, & Latham, 2016). However, the risks of forced-distribution systems outweigh their benefits for jobs that involve task interdependence and social support from others. Overall, Moon et al. (2016) recommended

using forced-distribution systems to rate a limited subset of activities—those that involve independent work effort and those that can be measured using objective performance measures.

Absolute Rating Systems

Absolute rating systems enable a rater to describe a ratee without making direct reference to other ratees.

Essays

Perhaps the simplest absolute rating system is the narrative essay, in which the rater is asked to describe, in writing, an individual's strengths, weaknesses, and potential and to make suggestions for improvement. The assumption underlying this approach is that a candid statement from a rater who is knowledgeable of a ratee's performance is just as valid as more formal and more complicated appraisal methods.

The major advantage of narrative essays (when they are done well) is that they can provide detailed feedback to ratees regarding their performance. Drawbacks are that essays are almost totally unstructured, and they vary widely in length and content. Comparisons across individuals, groups, or departments are virtually impossible, since different essays touch on different aspects of ratee performance or personal qualifications. Finally, essays provide only qualitative information; yet, for the appraisals to serve as criteria or to be compared objectively and ranked for the purpose of an employment decision, some form of rating that can be quantified is essential. Behavioral checklists provide one such scheme.

Behavioral Checklists

When using a behavioral checklist, the rater is provided with a series of descriptive statements of job-related behavior. His or her task is simply to indicate ("check") statements that describe the ratee in question. In this approach, raters are not so much evaluators as they are reporters of job behavior. Moreover, ratings that are descriptive are likely to be higher in reliability than ratings that are evaluative (Stockford & Bissell, 1949), and they reduce the cognitive demands placed on raters, valuably structuring their information processing (Hennessy, Mabey, & Warr, 1998).

o be sure, some job behaviors are more desirable than others; checklist items can, therefore, be scaled by using attitude-scale construction methods. In one such method, the Likert method of summated ratings, a declarative statement (e.g., "she follows through on her sales") is followed by several response categories, such as "always," "very often," "fairly often," "occasionally," and "never." The rater simply checks the response category he or she feels best describes the ratee. Each response category is weighted—for example, from 5 ("always") to 1 ("never") if the statement describes desirable behavior—or vice versa if the statement describes undesirable behavior. An overall numerical rating for each individual then can be derived by summing the weights of the responses that were checked for each item, and scores for each performance dimension can be obtained by using item analysis procedures (cf. Anastasi, 1988).

The selection of response categories for summated rating scales often is made arbitrarily, with equal intervals between scale points simply assumed. Scaled lists of adverbial modifiers of frequency and

amount are available, however, together with statistically optimal four- to nine-point scales (Bass, Cascio, & O'Connor, 1974). Scaled values also are available for categories of agreement, evaluation, and frequency (Spector, 1976).

Checklists are easy to use and understand, but it is sometimes difficult for a rater to give diagnostic feedback based on checklist ratings, for they are not cast in terms of specific behaviors. On balance, however, the many advantages of checklists probably account for their widespread popularity in organizations today.

Forced-Choice System

A special type of behavioral checklist is known as the forced-choice system—a technique developed specifically to reduce leniency errors and establish objective standards of comparison between individuals (Sisson, 1948). To accomplish this, checklist statements are arranged in groups, from which the rater chooses statements that are most or least descriptive of the ratee. An overall rating (score) for each individual is then derived by applying a special scoring key to the rater descriptions.

Forced-choice scales are constructed according to two statistical properties of the checklist items: (1) discriminability, a measure of the degree to which an item differentiates effective from ineffective workers, and (2) preference, an index of the degree to which the quality expressed in an item is valued by (i.e., is socially desirable to) people. The rationale of the forced-choice system requires that items be paired so that they appear equally attractive (socially desirable) to the rater. Theoretically, then, the selection of any single item in a pair should be based solely on the item's discriminating power, not on its social desirability.

As an example, consider the following pair of items:

Separates opinion from fact in written reports.

Includes only relevant information in written reports.

Both statements are approximately equal in preference value, but only item 1 was found to discriminate effective from ineffective performers in a police department. This is the defining characteristic of the forced-choice technique: Not all equally attractive behavioral statements are equally valid.

The main advantage claimed for forced-choice scales is that a rater cannot distort a person's ratings higher or lower than is warranted, since he or she has no way of knowing which statements to check in order to do so. Hence, leniency should theoretically be reduced. Their major disadvantage is rater resistance. Since control is removed from the rater, he or she cannot be sure just how the subordinate was rated. Finally, forced-choice forms are of little use (and may even have a negative effect) in performance appraisal interviews, for the rater is unaware of the scale values of the items he or she chooses. Since rater cooperation and acceptability are crucial determinants of the success of any performance management system, forced-choice systems tend to be unpopular choices in many organizations

Critical Incidents

This performance measurement method has generated a great deal of interest and several variations of the basic idea are currently in use. As described by Flanagan (1954a), the critical requirements of a job are those behaviors that make a crucial difference between doing a job effectively and doing it ineffectively. Critical incidents are simply reports by knowledgeable observers of things employees did that were especially effective or ineffective in accomplishing parts of their jobs. Supervisors record critical incidents for each employee as they occur. Thus, they provide a behaviorally based starting point for appraising performance. For example, in observing a police officer chasing an armed robbery suspect down a busy street, a supervisor recorded the following:

June 22, officer Mitchell withheld fire in a situation calling for the use of weapons where gunfire would endanger innocent bystanders.

These little anecdotes force attention on the situational determinants of job behavior and on ways of doing a job successfully that may be unique to the person described. The critical incidents method looks like a natural for performance management interviews because supervisors can focus on actual job behavior rather than on vaguely defined traits. Ratees receive meaningful feedback to which they can relate in a direct and concrete manner, and they can see what changes in their job behavior will be necessary in order for them to improve. In addition, when a large number of critical incidents are collected, abstracted, and categorized, they can provide a rich storehouse of information about job and organizational problems in general and are particularly well suited for establishing objectives for training programs (Flanagan & Burns, 1955).

As with other approaches to performance appraisal, the critical incidents method also has drawbacks. First, it is time consuming and burdensome for supervisors to record incidents for all of their subordinates on a daily or even weekly basis. Feedback may, therefore, be delayed. Nevertheless, incidents recorded in diaries allow raters to impose organization on unorganized information (DeNisi, Robbins, & Cafferty, 1989). Second, in their narrative form, incidents do not readily lend themselves to quantification, which, as we noted earlier, poses problems in between-individual and between-group comparisons, as well as in statistical analyses. For these reasons, a modification has been the development of behaviorally anchored rating scales, an approach we consider shortly.

Graphic Rating Scales

Probably the most widely used method of performance rating is the graphic rating scale, examples of which are presented in Figure 5.2. In terms of the amount of structure provided, the scales differ in three ways: (1) the degree to which the meaning of the response categories is defined, (2) the degree to which the individual who is interpreting the ratings (e.g., an HR manager or researcher) can tell clearly what response was intended, and (3) the degree to which the performance dimension being rated is defined for the rater.

Figure 5.2 Examples of Graphic Rating Scales

On a graphic rating scale, each point is defined on a continuum. Hence, to make meaningful distinctions in performance within dimensions, scale points must be defined unambiguously for the rater. This

process is called anchoring. Scale (a) uses qualitative end anchors only. Scales (b) and (e) include numerical and verbal anchors, while scales (c), (d), and (f) use verbal anchors only. These anchors are almost worthless, however, since what constitutes high and low quality or "outstanding" and "unsatisfactory" is left completely up to the rater. A "commendable" for one rater may be only a "competent" for another. Scale (e) is better, for the numerical anchors are described in terms of what "quality" means in that context.

The scales also differ in terms of the relative ease with which a person interpreting the ratings can tell exactly what response was intended by the rater. In scale (a), for example, the particular value that the rater had in mind is a mystery. Scale (e) is less ambiguous in this respect.

Finally, the scales differ in terms of the clarity of the definition of the performance dimension in question. In terms of Figure 5.2, what does quality mean? Is quality for a nurse the same as quality for a cashier? Scales (a) and (c) offer almost no help in defining quality, scale (b) combines quantity and quality together into a single dimension (although typically they are independent), and scales (d) and (e) define quality in different terms altogether (thoroughness, dependability, and neatness versus accuracy, effectiveness, and freedom from error). Scale (f) is an improvement in the sense that, although quality is taken to represent accuracy, effectiveness, initiative, and neatness (a combination of scale (d) and (e) definitions), at least separate ratings are required for each aspect of quality.

Graphic rating scales may not yield the depth of information that narrative essays or critical incidents do, but they (a) are less time consuming to develop and administer, (b) permit quantitative results to be determined, (c) promote consideration of more than one performance dimension, and (d) are standardized and, therefore, comparable across individuals. A drawback is that graphic rating scales give maximum control to the rater, thereby exercising no control over leniency, severity, central tendency, or halo. For this reason, they have been criticized. However, when simple graphic rating scales have been compared against more sophisticated forced-choice ratings, the graphic scales consistently proved just as reliable and valid (King, Hunter, & Schmidt, 1980) and were more acceptable to raters (Bernardin & Beatty, 1991).

Behaviorally Anchored Rating Scales

How can graphic rating scales be improved? According to Smith and Kendall (1963):

Better ratings can be obtained, in our opinion, not by trying to trick the rater (as in forced-choice scales) but by helping him to rate. We should ask him questions which he can honestly answer about behaviors which he can observe. We should reassure him that his answers will not be misinterpreted, and we should provide a basis by which he and others can check his answers. (p. 151)

Their procedure is as follows. At an initial conference, a group of workers and/or supervisors attempts to identify and define all of the important dimensions of effective performance for a particular job. A second group then generates, for each dimension, critical incidents illustrating effective, average, and ineffective performance. A third group is then given a list of dimensions and their definitions, along with a randomized list of the critical incidents generated by the second group. Their task is to sort or locate incidents into the dimensions they best represent (Hauenstein, Brown, & Sinclair, 2010).

This procedure is known as retranslation, since it resembles the quality control check used to ensure the adequacy of translations from one language into another. Material is translated into a foreign language by one translator and then retranslated back into the original by an independent translator. In the context of performance appraisal, this procedure ensures that the meanings of both the job dimensions and the behavioral incidents chosen to illustrate them are specific and clear. Incidents are eliminated if there is not clear agreement among judges (usually 60–80%) regarding the dimension to which each incident belongs. Dimensions are eliminated if incidents are not allocated to them. Conversely, dimensions may be added if many incidents are allocated to the "other" category.

Each of the items within the dimensions that survived the retranslation procedure is then presented to a fourth group of judges, whose task is to place a scale value on each incident (e.g., in terms of a seven- or nine-point scale from "highly effective behavior" to "grossly ineffective behavior"). The end product looks like that in Figure 5.3.

Figure 5.3 Scaled Expectations Rating for the Effectiveness With Which the Department Manager Supervises His or Her Sales Personnel

Source: Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 57, 15–22. Copyright 1973 by the American Psychological Association.

As you can see, behaviorally anchored rating scales (BARS) development is a long, painstaking process that may require many individuals. Moreover, separate BARS must be developed for dissimilar jobs. Nevertheless, they are used quite frequently. For example, results of a survey involving hotels showed that about 40% used BARS (Woods, Sciarini, & Breiter, 1998).

How have BARS worked in practice? An enormous amount of research on BARS has been published (e.g., Maurer, 2002). At the risk of oversimplification, major known effects of BARS are summarized in Table 5.3 (cf. Bernardin & Beatty, 1991). A perusal of this table suggests that little empirical evidence supports the superiority of BARS over other performance rating systems.

Summary Comments on Rating Formats and Rating Process

For several million workers today, especially those in the insurance, communications, transportation, and banking industries, being monitored on the job by a computer is a fact of life (Kurtzberg, Naquin, & Belkin, 2005; Tomczak, Lanzo, & Aguinis, 2018). In most jobs, though, human judgment about individual job performance is inevitable, no matter what format is used. This is the major problem with all formats.

Unless observation of ratees is extensive and representative, it is not possible for judgments to represent a ratee's true performance. Since the rater must often make inferences about performance, the appraisal is subject to all the biases that have been linked to rating scales. Raters are free to distort

their appraisals to suit their purposes. This can undo all of the painstaking work that went into scale development and probably explains why no single rating format has been shown to be superior to others.

What can be done? Both Banks and Roberson (1985) and Härtel (1993) suggest two strategies: First, build in as much structure as possible in order to minimize the amount of discretion exercised by a rater. For example, use job analysis to specify what is really relevant to effective job performance, and use critical incidents to specify levels of performance effectiveness in terms of actual job behavior. Second, don't require raters to make judgments that they are not competent to make; don't tax their abilities beyond what they can do accurately. For example, for formats that require judgments of frequency, make sure that raters have had sufficient opportunity to observe ratees so that their judgments are accurate.

As we discussed earlier, performance appraisal is a complex process that may be affected by many factors, including organizational, political, and interpersonal barriers. In fact, idiosyncratic variance (i.e., variance due to the rater) has been found to be a larger component of variance in performance ratings than the variance attributable to actual ratee performance (Greguras & Robie, 1998; Scullen, Mount, & Goff, 2000). For example, rater variance was found to be 1.21 times larger than ratee variance for supervisory ratings, 2.08 times larger for peer ratings, and 1.86 times larger for subordinate ratings (Scullen et al., 2000). In addition, raters may be motivated to inflate ratings for reasons completely unrelated to the true nature of an employee's performance, such as the desire to avoid confrontation with subordinates, promote a problem employee out of the unit, look like a competent manager, and procure resources (Spence & Keeping, 2011). In this section, we consider individual differences in raters and in ratees (and their interaction) and how these factors affect performance ratings. Findings in each of these areas are summarized in Tables 5.4, 5.5, and 5.6. For each variable listed in the tables, an illustrative reference is provided for those who wish to find more specific information.

As the tables demonstrate, we now know a great deal about the effects of selected individual differences variables on ratings of job performance. However, there is a great deal more that we do not know. Accordingly, there is ongoing research proposing new formats and procedures (e.g., Hoffman et al., 2012). Above all, however, recognize that the process of performance appraisal, including the social and emotional context, and not just the mechanics of collecting performance data, determines the overall effectiveness of this essential component of all performance management systems (Djurdjevic & Wheeler, 2014).

Evaluating the Performance of Teams

Our discussion thus far has focused on the measurement of employees working independently and not in groups. We have been focusing on the assessment and improvement of individual performance. However, numerous organizations are structured around teams (Hollenbeck, Beersma, & Schouten, 2012). Team-based organizations do not necessarily outperform organizations that are not structured around teams (Hackman, 1998). However, the interest in, and implementation of, team-based structures does not seem to be subsiding; on the contrary, there seems to be an increased interest in organizing how work is done around teams (Naquin & Tynan, 2003). Therefore, given the popularity of teams, it makes sense for performance management systems to target not only individual performance

but also an individual's contribution to the performance of his or her team(s), as well as the performance of teams as a whole (Aguinis, Gottfredson, & Joo, 2013a; Li, Zheng, Harris, Liu, & Kirkman, 2016).

The assessment of team performance does not imply that individual contributions should be ignored. On the contrary, if individual performance is not assessed and recognized, social loafing may occur (Scott & Einstein, 2001). Even worse, when other team members see there is a "free rider," they are likely to withdraw their effort in support of team performance (Heneman & von Hippel, 1995). Assessing overall team performance based on team-based processes and team-based results should therefore be seen as complementary to the assessment and recognition of (a) individual performance (as we have discussed thus far), and (b) individuals' behaviors and skills that contribute to team performance (e.g., self-management, communication, decision making, collaboration; Aguinis, Gottfredson, & Joo, 2013a).

Meta-analysis results provide evidence to support the need to assess and reward both individual and team performance because they have complementary effects (Garbers & Konradt, 2014). Thus, the average effect size of using individual incentives on individual performance is $g = 0.32$ (based on 116 separate studies), and the average effect size of using team incentives on team performance is $g = 0.34$ (based on 30 studies). (The effect size g is similar to d—a standardized mean difference between two groups.)

Not all teams are created equally, however. Different types of teams require different emphases on performance measurement at the individual and team levels. Depending on the complexity of the task (from routine to nonroutine) and the membership configuration (from static to dynamic), we can identify three different types of teams (Scott & Einstein, 2001):

Work or service teams: Intact teams engaged in routine tasks (e.g., manufacturing or service tasks)

Project teams: Teams assembled for a specific purpose and expected to disband once their task is complete; their tasks are outside the core production or service of the organization and, therefore, less routine than those of work or service teams.

Network teams: Teams whose membership is not constrained by time or space or limited by organizational boundaries (i.e., they are typically geographically dispersed and stay in touch via telecommunications technology); their work is extremely nonroutine.