Murphy Ch 5

5 OBTAINING INFORMATION AND EVALUATING PERFORMANCE

Learning Objectives

- 5.1 Understand the advantages and disadvantages of using peers, self-ratings, and subordinate ratings versus or as adjuncts to ratings from direct supervisors and managers
- 5.2 Learn how electronic performance monitoring is used in organizations and why it is not always a good substitute for judgmental measures of performance
- 5.3 Understand the roles of attention and memory in determining the performance judgments of supervisors and managers
- 5.4 Learn how liking and affect can influence performance ratings
- 5.5 Understand the processes used to set performance goals and standards

Performance appraisal requires people to make judgments about job performance based on observations they have made, knowledge they have obtained, and information they might receive. In this chapter, we consider two important facets of this process: (1) *who*—who should be charged with the task of evaluating job performance, and what are the advantages and disadvantages of different options an organization might consider? and (2) *how*—how does that person or set of persons acquire and make sense of the information needed to make these judgments? The *who* question is strongly related to a number of factors discussed in previous chapters. For instance, the question of who is given the responsibility of evaluating performance might depend on cultural factors (both national and organizational cultures) as well as the purpose of performance appraisal (different evaluators might make sense, depending on the purpose of performance appraisal in a particular organization).

Traditionally, the question of *who* completes your performance appraisal has rarely been an issue. In most organizations, appraisals are completed by one's immediate supervisor or manager, and each supervisor or manager in an organization has been solely responsible for evaluating his or her direct reports. Supervisory evaluations are still very important, and it is the norm in virtually all organizations that use performance appraisal to collect evaluations from supervisors, but there are many other options. It might be feasible to obtain evaluations of performance from self-ratings, from peers, from subordinates or customers, or some combination of any of these. Indeed, many organizations use some sort of multisource evaluation or feedback system, and it is not unusual to collect performance evaluations and feedback from many sources (Morgeson, Mumford, & Campion, 2005).

The *how* issue is to some extent distinct from many of the issues that have been discussed in previous chapters, although the structure of workplaces and organizations can have a bearing on precisely how information about performance is obtained. For example, both the content of what people observe or consider, and the standards they use to evaluate that information, are substantially influenced by the context in which performance appraisal occurs. Nevertheless, there are some cognitive processes involved in forming judgments about job performance that are likely to operate similarly in a wide range of different contexts.

Who Should Evaluate Performance

In discussing the different players who might be involved in evaluating a person's performance, we find it useful to arrange this discussion in terms of the distance between the target and the evaluator. Thus, we will start by considering self-ratings, then move on to the rating source that is most proximal and most similar to the target – i.e., peers. We then consider supervisors, and finally other sources (e.g., subordinates, customers, some combination of multiple sources).

Self-Ratings

Many performance rating systems include self-ratings, and there is evidence that these ratings influence the judgments of supervisors; employees who give themselves higher self-ratings tend to receive higher ratings from their supervisors (Shore, Adams, & Tashchian, 1998). There are several reasons why it might be useful to include self-ratings as a part of performance appraisal. First, asking for self-ratings at the same time as you are asking supervisors to evaluate performance may help supervisors and subordinates adopt a common frame of reference when discussing job performance, particularly if both self-ratings and supervisory ratings are obtained on the same set of performance dimensions. By focusing both raters and rates on specific aspects of performance over the past year (or over some other period for organizations that obtain performance ratings on a different schedule), including self-ratings in the process might help raters and ratees reach some common understandings of what the ratee did well and what he or she did poorly during the rating period. Self-ratings also give employees some voice in the appraisal process, which can enhance perceptions of the fairness and reasonableness of the appraisal system (Korsgaard & Roberson, 1995; Lizzio, Wilson, & MacKay, 2008; Whiting, Podsakoff, & Pierce, 2008). That is, asking ratees to evaluate their own performance implies that the organization values their perspective and takes their input seriously. Unfortunately, including self-ratings in a performance appraisal system can be a source of conflict, because self-ratings and supervisory ratings often disagree. In particular, it is common to find that self-ratings are higher than supervisory ratings of performance (Harris & Schaubroeck, 1988; Meyer, 1980; Thornton, 1980). In fact, one source of resistance to including self-ratings in performance appraisal is the belief that they are often unrealistically high.

Are Self-Ratings Inflated?

The finding that self-ratings of job performance tend to be higher than supervisory ratings has been so widely replicated that the question is not *whether* self-ratings are likely to be higher than ratings from others, but rather *why* self-ratings are so high. One possibility is that inflated self-ratings are a result of *fundamental attribution error* (Jones & Nisbett, 1971; Ross, 1977) that is, the tendency to explain one's successes in terms of stable personal factors (e.g., ability, effort) and explain failures in terms of unstable situational factors (e.g., bad luck, someone else cheating), whereas others' successes are attributed to these situational factors, while their failures are attributed to lack of ability or lack of effort. This asymmetry in understanding why people succeed or fail will lead you to overestimate your own performance (because you see yourself as responsible for your success, but see situational explanations for your failures) and underestimate others' performance (e.g., believing others' success is due to luck or cheating). While this error is widely cited in social psychology, evidence for its effects is somewhat equivocal. On the whole, there is clearer evidence for biases in explanations for one's own success than for biases in explaining one's own failures (Miller & Ross, 1975). Nevertheless, there is a good deal of evidence that this bias in our perceptions of our own performance versus the performance of others will lead to systematic over-rating of one's own performance and systematic under-rating of others' performance (Campbell, Campbell, & Ho-Beng, 1998).

The fact that self-ratings are usually higher than supervisory ratings is not by itself necessarily evidence of the invalidity of self-ratings, since it is possible that self-ratings are accurate and supervisory ratings are unduly harsh. Similarly, evidence that self-ratings are not highly correlated with supervisory ratings (Landy & Farr, 1983; Steel & Ovalle, 1984; Thornton, 1980) is difficult to evaluate, because supervisory ratings are not the optimal criteria for evaluating self-assessments. However, there are two lines of evidence that would seem to point unambiguously to deficiencies in self-ratings. First, self-ratings tend to move closer to supervisory ratings if extensive performance feedback is given (Steel & Ovalle, 1984). Second, there is evidence that self-ratings are less lenient if raters know that the ratings will be checked against some objective criterion (Farh & Werbel, 1986). The finding that self-ratings tend to be higher than ratings from *any* other source supports the hypothesis that there are self-seeking biases in self-ratings of performance.

Heidemeier and Moser (2009) propose a different explanation for inflated self-ratings, noting that external observers have access to a different set of cues and they integrate them differently than individuals do when evaluating their own performance. That is, self-ratings might be based on a different set of information than ratings provided by others. This hypothesis is consistent with the findings of Williams and Johnson (2000), who report that self–supervisor agreement is higher in environments where there is more frequent feedback and more sharing of information between supervisors and subordinates, as well as the findings of Schrader and Steiner (1996), who found that self-ratings were more similar to supervisory ratings when common standards were used for evaluating behavior in the workplace. Finally, Brutus and Fleenor (1999) note that discrepancies between self-ratings and ratings provided by others vary by gender (female employees are less likely to inflate self-ratings), age (older employees are more likely to inflate), and organizational level (managers at higher levels in the organization are more likely to inflate self-ratings than lower-level managers).

Leniency bias in self-ratings is not necessarily universal. Farh, Dobbins, and Cheng (1991) suggested that the self-ratings may be *lower* than ratings from others in Asian cultures, and that in these cultures there might be a modesty bias. Subsequent studies have painted a more complicated picture. While there are some broad differences in the levels of individualism and collectivism (cultural dimensions assumed to be responsible for leniency and modesty biases) in Asian nations as compared to North American and European nations, there is significant nation-to-nation variability, and it is unwise to assume that there are distinct East–West differences in the structure or outcomes of performance appraisal (Barron & Sackett, 2008; Yu & Murphy, 1993). Thinking about self-ratings in particular, it appears that self-ratings are sometimes higher than and sometimes lower than ratings from others in Asian nations, and that nation-specific cultural factors may be as important as broad East–West cultural differences for understanding the conditions under which self-ratings will or will not be higher than ratings from others.

In addition to cultural factors, there are characteristics of the task and the rating environment that can lead to people to under-rate their own performance. In particular, when engaging in group tasks, there is evidence that self-ratings are a bit lower than they should be, while ratings of the group's performance as a whole tend to be slightly inflated (Jourden & Heath, 1996). This tendency may be an instance of group ethnocentrism, because people consistently rate their own group as performing better than other peoples' groups. Nevertheless, this finding does suggest that, even in Western cultures, inflation of self-ratings is not a uniquely self-serving process, and that this inflation extends to one's own group, perhaps even more strongly than to self-ratings.

Do Self-Ratings Contribute to the Accuracy or Acceptance of Appraisal?

In principle, including self-ratings as part of a performance appraisal system should increase ratees' satisfaction with the appraisal process because they have some voice in their appraisal and this increases the quality of the conversation between the rater and ratee. Unfortunately, it is not clear whether self-rating delivers these advantages. Roberson, Torkel, Korsgaard, Klein, Diddams, and Cayer (1993) conducted careful evaluation of the effects of including self-ratings, randomly assigning groups of ratees to rating conditions in which both the rater and ratee provided performance ratings or conditions in which there were no self-ratings. They found no evidence of systematic improvements, and in fact found that self-rating was associated with *lower* levels of perceived influence on the appraisal discussion and lower levels of rater–ratee agreement.

From the rater's perspective, self-ratings can be problematic, especially if the rater has access to the ratee's self-rating prior to completing the performance appraisal form. Shore et al. (1998) showed that raters who know that the self-rating is high are likely to inflate the ratings they give. This makes sense, because discrepancies between supervisory ratings and self-ratings are likely to lead to conflict and discomfort when giving performance feedback; by minimizing the spread between your own rating and your subordinate's self-rating, you can avoid this conflict.

Peer Ratings

Peer rating systems have long been used in the military (Landy & Farr, 1983), but until recently, they were more rarely used in other settings (Bernardin & Klatt, 1985; McEvoy & Buller, 1987). With the introduction of 360-degree feedback systems (Dierdorff & Surface, 2007; Levy & Williams, 2004; Morgeson et al., 2005; Seifert, Yukl, & McDonald, 2003), peer ratings of performance have become substantially more common.

Peiperi (1999) noted that the success or failure of peer rating systems probably depends on a number of factors, ranging from the organizational culture and the types of positions involved to the purpose of rating and the methods used to decide who rates who. He tested an elaborate model of the likely causes and indicators of the success or failure of peer rating systems, and found strong support for some predictions (e.g., the interdependence of work and the use of performance-based rewards is related to successful peer rating), with mixed support for others (e.g., he predicted that peer ratings would be successful in cohesive cultures, but found that cohesion actually *reduces* the success of peer rating, perhaps because cohesive groups resist rating systems that attempt to distinguish performance levels within groups). Other researchers have examined more narrow hypotheses about the determinants of peer rating success. For example, Maurer & Tarulli, 1996) report that peer ratings are more likely to be accepted if it is believed that peers have the opportunity to observe relevant behaviors.

There are some clear advantages to obtaining performance ratings from peers. Because peers often work in close proximity, they have ample opportunity to observe each other's behavior. Peers' frequent opportunity to observe task behaviors, interpersonal behaviors, and results may make them a uniquely valuable source. Wherry's theory of rating cites rater–ratee proximity as one key to accuracy in rating (Wherry & Bartlett, 1982), and this theory suggests that the opinions of peers are especially reliable because they have more opportunity than supervisors to observe each other. However, Imada (1982) cautions that there is little data to support the assumption that peers' opportunity to observe behavior is related to the accuracy of their judgments. Considerations of the work roles of peers might even suggest that their observations will *not* be the best source of information about a particular individual's performance. Unlike supervisors, evaluation of the performance of other employees is not a core part of the job, and peers may be less systematic in their observation or in their evaluation of what they observe.

Despite the shortcomings of peer ratings noted above, there is evidence for the reliability and validity of peer ratings (Gregarus & Robie, 1998; Saavedra & Kwun, 1993), and there are many reasons to believe that incorporating information from peers could improve the quality of performance appraisal. Peers often see different aspects of performance than supervisors see (especially citizenship behaviors, many of which are directed at peers), and the perspective they apply in evaluating those behaviors may be more similar to that of the ratee than the evaluative framework applied by supervisors.

There is evidence that peers anchor their ratings in terms of their own performance levels. That is, higher-performing peers have more demanding standards when evaluating others' performance (Saavedra & Kwun, 1993). There is also evidence that, like other raters, peers' judgments about the performance of their colleagues are influenced by the degree to which they like each other or have similar personalities (Antonioni & Park, 2001).

The behavior observed by peers is both quantitatively and qualitatively different from that observed by supervisors and subordinates, in that peers see both more behaviors and different behaviors than do other sources. This is particularly true in the domain of interpersonal relations. An individual may be on his or her guard when interacting with superiors or subordinates, but is more likely to behave naturally (whatever the person's natural style of interaction) among peers. Peers are also likely to encounter secondhand information about interpersonal behaviors; colleagues tend to talk about colleagues, and interpersonal issues are likely to be a frequent topic of conversation.

Resistance to Peer Rating

Despite the advantages noted above, there are several sources of resistance to the use of peer ratings in evaluating performance. First, the use of peer ratings violates the traditional power hierarchy in organizations. Traditionally, evaluations have flowed in a top-down fashion, in which people at higher authority levels evaluate the performance of their subordinates. Peer rating violates this principle. Second, the use of peer ratings puts workers in an uncomfortable position that could interfere with their ability to work together effectively. By forcing people who work side by side and who may work together on joint projects to evaluate one another, appraisal systems that include peer ratings will require workers to make hard decisions about their colleagues, increasing the possibility that negative peer feedback will make it difficult for the work group to work together well in the future. Because peers are often directly competing for the same rewards (e.g., it is common for organizations to set aside a specific amount of money for raises), peer rating systems can also introduce unhealthy competition and self-seeking behavior. That is, a person who rates his or her peers as poor performers may increase the probability that *he or she* will receive rewards that peers are denied because of the low ratings they receive. Research on peer ratings suggests that these ratings are more readily accepted if they are used for developmental purposes than if they are used to help determine pay raises or promotions (McEvoy & Buller, 1987).

The resistance to peer ratings does not seem to be connected to concerns over their psychometric shortcomings. There is ample evidence that peer ratings show levels of reliability and validity comparable to supervisory ratings (Kane & Lawler, 1978, 1980; Landy & Farr, 1983; Wexley & Klimoski, 1984). An advantage of peer ratings is that they can be pooled, a procedure that can substantially increase reliability and partially remove idiosyncratic biases of any particular rater (Kenny & Berman, 1980; Murphy, 1982a). Despite these potential advantages to including peers in performance rating, there is often substantial resistance to this idea. It appears that the resistance to peer ratings can be traced to two causes: (a) concerns over role reversals, and (b) concerns over distortion. The first issue, role reversals, has been noted earlier. Incorporation of peer ratings changes the power structure of an organization in such a way that people who are at the same level in the formal power hierarchy nevertheless have power over one another.

There are two reasons that lead many managers to conclude that peer ratings will be unduly lenient. First, it is widely believed that the friendship between peers will lead to inflated ratings (Landy & Farr, 1983). In fact, there is little evidence that friendship bias is an important factor, although bias might be a problem if the ratee is aware of individual raters' scores. A second problem, and a potentially more serious one, is range restriction. It is plausible that peers will be unwilling to differentiate good from poor performers in the work group, for fear of "rocking the boat." While this problem is no doubt present in some settings, there is little evidence to suggest that peers are any more susceptible to range restriction than are supervisors.

Supervisors as Raters

Mayhew (1983) surveyed societies and organizations ranging in time from the sixth century B.C. to the present, and in scope from the Achaemenid Empire of ancient Mesopotamia and Persia to the modern police departments, and concluded that hierarchical differentiation with a unity of the flow of authority and power from the top down is essentially universal. That is, organizations are structured so that decisions, orders, and control flow from the top levels down, with very few instances of egalitarian or bottom-up rule. One of the many manifestations of hierarchical power relations in organizations is that evaluations usually flow from higher levels to lower ones. Scott (1975) noted, “evaluation is required if power is to be employed to control behavior” (p. 134). Thus, one argument for obtaining performance evaluations from supervisors or direct managers is that it is the normal thing to do in an organization that is hierarchically structured.

Evaluation is not just a social norm; it is tied directly to the nature and key requirements of the job of supervisor or manager. A supervisor has the responsibility of making sure that his or her subordinates complete assigned tasks within the constraints of resources and budgets, so it is critical that he or she is aware of how well each subordinate is performing. Indeed, because of this responsibility for ensuring that key production or performance goals are met, a person’s supervisor or direct manager is often best informed about the demands and constraints of a job and of each subordinate’s success in meeting the key objectives of that job. In addition, supervisors and direct managers often have considerable direct knowledge about the job. Supervisors, in particular, are often promoted from the ranks of workers in the job they now supervise, and are thus genuine subject matter experts. It is little surprise, therefore, that input from supervisors or direct managers is an almost universal component of performance appraisal systems in organizations.

Changes in the workplace are slowly disrupting the normal flow of evaluations from the direct supervisor to the ratee. As Thomas (1999) notes, it is increasingly common to find that workers have no immediate supervisor. Consider, for example, telecommuters who work from home. There may be supervisory or managerial employees who are familiar with the *results* of their work, but it is unlikely that there is any supervisor who is familiar with the work itself. This raises the question of who (if anyone) is sufficiently qualified and sufficiently familiar with the work being performed to complete performance appraisals. To the extent that work is performed with little meaningful supervision, it is possible that performance appraisal will simply not be possible in any meaningful sense. If there is no supervisor or direct manager with knowledge about the work behavior of a particular employee, it is unlikely that there will be others in the organization who have that knowledge. It will still be possible to measure the results of performance, but if the work is done remotely, it might not be possible to measure the behaviors that constitute job performance.

Using Ratings From Subordinates in Performance Appraisal

Upward appraisals, where subordinates evaluate their superiors, are rare (Hall, Leidecher, & DiMarco, 1996), except perhaps as part of a broader 360-degree evaluation system. In many ways this is a shame, because there is evidence that upward appraisals can improve both communication and employee satisfaction. Subordinates have a unique perspective and are likely to observe and experience aspects of their supervisor’s performance that are not readily observed by other potential raters. Of course, these systems can be uncomfortable for raters, who may reasonably fear that giving poor performance feedback to their supervisor will negatively influence *their* evaluations. For this reason, Hall et al. (1996) suggest that upward feedback be anonymous (e.g., supervisors could receive information about the distribution of ratings they received, without necessarily being aware of who gave what ratings). Nevertheless, the potential value of these systems would seem to suggest that they should be more common.

First, each potential rating source sees different behaviors and different outcomes, and subordinates may be able to add unique information to the performance evaluation process. Second, the use of subordinates provides an opportunity (also present in peer ratings, though perhaps to a less extent) to increase the reliability of information about performance by aggregating information from several raters. Third, the use of subordinate input in a performance appraisal system conveys a potentially powerful message about the climate and culture of the organization (i.e., its willingness to buck traditional norms), and it might even be thought of as an intervention used to influence the culture of an organization. That is, if you wanted to install a more egalitarian culture, one place to start might be to solicit input from subordinates when evaluating the performance or their supervisors and managers.

In our view, the biggest obstacle to implementing upward feedback systems is similar to the main obstacle to the use of peer ratings—it involves considerations of power and influence. Involving subordinates in evaluation is an even more serious violation of the status hierarchy than involving peers (Dornbusch & Scott, 1975; Thompson, 1967).

The typical definition of a supervisor in an organization is as someone who has the right and responsibility to evaluate the performance of some number of employees who report to him or her. Upward feedback violates this fundamental dynamic, upending the power hierarchy in organizations. The rigid distinction between management and labor, or between supervisors and subordinates, is violated if subordinates evaluate their superiors *and* those evaluations are taken seriously. Some organizations might be happy to upend this hierarchy, but hierarchical organization is still certainly the norm, and upward feedback may often turn out to be uncomfortable to both raters and ratees.

Spotlight 5.1 Are You a Biased Rater?

One of the common concerns in traditional performance appraisal systems is that rater biases can influence performance ratings. Raters might be biased against members of particular groups (e.g., racial and ethnic minorities, older workers, women) or they might simply be biased in favor of or against particular individuals who they like or dislike. While many people think their manager or supervisor is a biased rater, few people think they themselves are biased. However, with the widespread availability of the Implicit Association Test (IAT), peoples' beliefs about their ability to give ratings that are unbiased or fair is increasingly being challenged.

The IAT is a computerized test that purports to reveal peoples' biases, even ones people are completely unaware of (Greenwald, McGhee, & Schwartz, 1998). There are many variations of the IAT (Harvard's Project Implicit offers tests assessing biases involving sexuality, race, skin tone, weight, age, religions, and many other categories), but in general, they involve computerized tasks that require you to associate pictures with concepts, as quickly as possible. For example, you might be presented with a picture of a face and classify it, as quickly as you can, with "Old" or "Young" categories. There are many different tasks that are used in IATs, but the general idea is that small but consistent differences in the speed of association and in the links people make between pictures and categories reveal hidden biases. It is not unusual for people to take a test of this sort and come away saying, “I never realized I was biased against….”

There are significant questions about the meaning and validity of IAT scores (Greenwald, Banaji, & Nosek, 2015; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). For example, suppose you found, using an IAT, that a particular rater was biased against people of different races. Does that mean that he or she would be likely to give lower ratings to minority group members he or she supervises? The best answer is that nobody knows. There are often significant problems using results from social psychology experiments to predict important outcomes in organizations (Copus, Uglow, & Sohn, 2005), and it is not clear whether differences of a few milliseconds in the speed of associating pictures to concepts will tell you much about how a supervisor will evaluate people who he observes and works with on a day-to-day basis.

It is a good bet that if you take the IAT, you will be shown to exhibit biases toward or against some groups, possibly toward or against many groups. Does this mean you cannot give people a fair shake when you evaluate their performance? In our view, it is a long and risky leap from current tests of implicit biases to drawing meaningful conclusions about rating behavior.

Electronic Performance Monitoring

Traditional discussions of performance appraisal have been built around a model where some rater or group of raters observes or obtains information about the performance of a set of employees and makes judgments about their performance. Electronic performance monitoring (EPM) systems represent application of electronic technology to observe, record, and monitor various aspects of employee performance (Bhave, 2014), and in principle it might be possible to bypass subjective measures such as ratings altogether if sufficient data about performance can be obtained electronically. For example, Neary (2002) describes the development of a computerized performance management system for nearly 100,000 employees at TRW Automotive.

Organizations frequently use a range of electronic monitoring methods to collect information about the performance and behaviors of employees (Wells, Moorman, & Werner, 2007). These range from monitoring web usage to keystroke monitoring; some organizations even use GPS data to monitor the whereabouts and movements of their employees. Reactions to this electronic monitoring are often negative, but this might depend on the perceived purpose of this monitoring. Wells et al. (2007) suggest that if the purpose of this monitoring is seen as developmental, reactions may be more positive than if the purpose of monitoring is seen as a technique for controlling employee behavior.

Electronic monitoring systems have some clear advantages over reliance on human observers. EPM systems allow continuous monitoring, and they provide unbiased information about employee behavior and performance. There is evidence that the use of EPM systems as one tool in performance appraisal (in particular, in appraisal systems where the supervisor receives EPM reports and considers them when evaluating subordinates) is associated with increases in task performance (Bhave, 2014; Goomas, 2007; Kolb & Aiello, 1997). These systems provide both supervisors and subordinates with detailed information about the performance of many aspects of the job, and provide an opportunity to obtain frequent, verifiable, and credible feedback (Goomas, 2007).

There are a number of potential limitations to EPM systems. First, they place a great premium on aspects of job performance that are easily measured. For example, Goomas (2007) describes an EPM system in an automotive parts warehouse that included measurement of travel time (seconds per foot) and time spent handling each case of material; to evaluate performance, these metrics were compared to standard metrics established through careful industrial engineering studies. The problem with focusing on behaviors and outcomes that are most easily countable is that you might end up missing critically important aspects of task performance, such as exercising judgment, planning, communication, as well as missing just about all aspects of citizenship and adaptive performance. An overemphasis on EPM is likely to lead to criterion deficiency, overemphasizing the easily countable aspects of performance and underemphasizing the rest.

Employees whose work is electronically monitored (for example, call center employees whose keystrokes and call times are often recorded and fed back to supervisors) often find such systems intrusive and stressful (Bates & Holton, 1995; Sewell, Barker, & Nyberg, 2011), particularly if these systems are seen as a tool for speeding up and intensifying work. There is empirical evidence that EPM systems can increase stress (see, for example, Aiello & Kolb, 1995), and it is possible that the productivity gains that can be achieved by speeding up work (a common outcome of EPM) will be offset by the increased costs due to stress-related health problems among workers. Even if the organization does not suffer ill effects resulting from the stresses it imposes on employees, the employees themselves are quite likely to suffer if these systems are put into place. The stressful nature of EPM can be mitigated somewhat if employees are given a role in the development and use of EPM.

Non-Electronic Alternatives

Stanton (2000) notes that performance monitoring existed well before electronic methods of monitoring were available, and that non-electronic monitoring continues to this day. Human judges can be used to collect and categorize detailed records about performance, and despite some clear differences between EPM and the use of human monitors (e.g., EPM is likely to be continuous, and is not hampered by the perceptual and cognitive limitations of human monitors), there are clear similarities. Both humans and electronic devices monitor performance for reasons that range from providing feedback to motivating employees to perform at a higher level. Both can be stressful and can contribute to negative perceptions of the job and the organization. Stanton (2000) suggests that trust in management and trust in supervisors is a critical factor in determining reactions to performance monitoring, and that when trust is low, monitoring is more likely to be seen as punitive and coercive.

Agreement Across Rating Sources

Self-ratings, peer ratings, supervisory ratings, ratings from subordinates, and other sources of information for evaluating performance often lead to somewhat different conclusions about the performance and about the strengths and weaknesses of the people being evaluated. You could think of this as either an advantage or as a disadvantage of using multiple sources in evaluating performance. This disagreement is a good thing in the sense that is suggests that each rating source might have something unique to contribute. If self-ratings, peer ratings, supervisory ratings, and others all agreed, there would hardly be any point in collecting information from multiple sources. On the other hand, disagreement can undercut the credibility of the appraisal system. We noted earlier that self-ratings tend to be high, and that people often rate their own performance higher than it would be rated by others. If ratings from multiple sources disagree, this implies that the rating a particular employee receives from some sources will be higher than the ratings he or she receives from other sources. For example, suppose your peers rate you a “4” on oral communication, but your supervisor and your subordinates give you ratings of “2” and “3,” respectively. If, like most people, you over-rate your own performance, you are likely to accept the peer rating as most accurate, *and* to use the fact that some raters gave you a “4” to cast doubt on the accuracy of the “2” and “3” ratings you receive. Thus, disagreements between raters might both cast some general doubt on the accuracy of evaluations and provide evidence that the ratee can use to discredit whoever gives him or her a low rating.

There have been several reviews of research on agreement between ratings obtained from different sources, and in general they suggest that different ratings sources tend to show low to moderate levels of agreement. (Greguras & Robie, 1998; Harris & Schaubroeck, 1988). For example, correlations between supervisor ratings and peer ratings of performance tend to be in the .30–.40 range (Conway & Huffcutt, 1997; Viswesvaran, Schmidt, & Ones, 2002). Self-ratings and ratings from other sources do show even lower levels of agreement (Atwater, Ostroff, Yammarino, & Fleenor, 1998; Landy & Farr, 1983; Steel & Ovalle, 1984; Thornton, 1980). For example, the correlations between self-ratings and peer ratings and between self-ratings and supervisory ratings tend to be closer to .20 (Conway & Huffcutt, 1997; Heidemeir & Moser, 2009). Within-source agreement (e.g., the correlation between ratings given by two different peers) is typically higher than between-source agreement (Conway & Huffcutt, 1997), which suggests that the unique perspectives and the differences in what peers, supervisors, subordinates, and others tend to observe contributes to the relatively low level of agreement. Both demographic (e.g., gender, age) and personality factors (e.g., dominance, empathy) are related to cross-level agreement (Brutus, Fleenor, & McCauley, 1999).

How the Structure of Organizations and Workgroups Influences Observation of Work Behavior and Agreement Regarding Performance

The changing nature of the workplace means that supervisors are unlikely to have the opportunity to directly observe the performance of their subordinates all of the time. Approximately 20–25% of the U.S. workforce telecommutes at least some of the time, and approximately 50–60% of workers in Fortune 1000 companies work away from their desks at least some of the time (Latest Telecommuting Statistics, 2015). As a result, it is likely that at least some of the information supervisors receive about the performance of their subordinates will be indirect, either secondhand reports from other sources or inferences the supervisor might make from the work products he or she sees. We should note that even in traditional workplaces, supervisors often obtain information about the performance of their subordinates from sources other than direct observation (Raymark, Balzer, & DeLaTorre, 1999); however, with the growth of telecommuting and other methods of working away from the traditional workplace, the use of indirect information in evaluating performance is likely to become more common and more important.

The question naturally arises of how supervisors put together whatever direct information they receive by observing subordinates with indirect information to arrive at an overall evaluation of each subordinate’s performance. There is evidence that supervisors pay considerably more information to performance they have observed than they do to indirect information about the performance of their subordinates (Golden, Barnes-Farrell, & Mascharka, 2009). Indirect information that is inconsistent with the rater’s own observations receives little weight in evaluating performance (Uggerslev & Sulsky, 2002).

Although there are few empirical tests of this hypothesis, we believe that differences in what people at different levels of an organization observe are at least in part responsible for differences in performance ratings. These differences in observation are driven by a wide range of factors, from the physical arrangement of the workplace to the likelihood that particular individuals will either work together or work in settings where they can observe coworkers’ performance. It is also likely that changes in the structure of workplaces will have a strong influence on across-source differences in opportunities to observe work behavior. Many organizations are moving in the direction of an “open office,” where employees share work spaces (e.g., multiple employees might work at the same table and few employees might have their own office). This trend is not universally welcomed (Kaufman, 2014), and its implications for the structure of work are still unfolding, but it is likely to influence the dynamics of performance appraisal. In an open office, more individuals will have an opportunity to observe more of your work behavior, and it is likely that these observations will filter into appraisals.

Cognitive Processes in Performance Evaluation

Research on performance appraisal and the potential difficulties in forming sound judgments about performance can be traced back almost 100 years (Thorndike, 1920), and until the late 1970s much of this work proceeded on the assumption that supervisors and managers observed the behavior of those they are asked to rate with reasonable accuracy, and that at the time they were asked to make judgments about performance, could recall a good deal of this information and pull it together to make judgments about performance. Thus, the job of industrial and organizational psychologists working in this field was to give raters the tools (e.g., well-developed rating scales, rater training) to help them do the best possible job evaluating their subordinates’ performance. By the 1970s and 1980s, the cognitive revolution in psychology had completely overturned the assumption that human cognition worked like a movie camera, faithfully recording what was seen and allowing you to replay it with reasonable accuracy at some later time. Rather, the set of processes involved in acquiring and making sense of information about the external world was understood to be a complex one that involved multiple systems for processing information (Kahneman, 2011; Shiffrin & Scheider, 1977), in which categories and prototypes often structure the storage and retrieval of information (Feldman, 1981) and in which even the most vivid and concrete memories might be false or misleading (Loftus, 2005).[1]

Landy and Farr's (1980) review of research on performance appraisal suggested that interventions such as improving rating scales had little impact on the quality of rating data, and this review was instrumental in turning the field toward a more careful examining of the cognitive processes involved in performance rating. Several subsequent authors (e.g., Murphy & Cleveland, 1991; DeNisi, 2006) noted that cognitive research is probably more relevant to understanding *performance judgment* than understand the ratings that are given in organizations, noting that ratings that are actually recorded on performance appraisal forms do not always correspond to the rater's judgments regarding the performance level or the strengths and weaknesses of the individuals being rated. Nevertheless, research that gives us a better understanding of how judgments about performance are formed has the potential to shed light on performance ratings, even when these ratings do not precisely mirror these performance judgments.

DeNisi, Cafferty, and Meglino (1984) developed an influential model of the cognitive processes underlying performance appraisal; the studies that led to this model and the development of the key concepts in this model are described in DeNisi (2006). This model and others developed at about the same time (e.g., Feldman, 1981) identified a set of activities a rater goes through when making judgments about a subordinate's performance. These are shown in Table 5.1. First, the rater must observe and form a mental representation of behaviors that are part of the domain of job performance. Next, these representations are stored in memory, and at some later time retrieved. Next, information about performance, both what the rater has observed and other information obtained by inference (e.g., if the results of behavior on the job are favorable, the rater might infer that the behaviors themselves are appropriate) or from other sources (e.g., feedback from customers) must be integrated to form a judgment about the ratee's effectiveness.

Table 5.1 suggests a relatively straightforward linear process in which information is acquired, stored, retrieved, and integrated to form judgments. Research in the 1980s and 1990s suggested that this process was neither linear nor straightforward, and that this process could loop and bend in numerous ways, so that judgments made today might influence attention, encoding, and retrieval tomorrow, or that information that had been stored in memory in one way might be reprocessed and might, when retrieved, look quite different from what was originally stored. Although some of the most basic cognitive processes in forming judgments might be pretty much the same across a wide range of contexts, there were a number of contextual factors that have turned out to be very important for understanding how people obtain and process the information needed to make judgments about job performance.

Table 5.1 Cognitive Processes That Underlie Judgments About Job Performance

1. Behavior is observed by the rater
2. Rater forms a cognitive representation of the behavior
3. This representation is stored in memory
4. Stored information is retrieved from memory when needed
5. Retrieved information is integrated, along with other relevant information, to form a judgment about the ratee

Source: From DeNisi (2006), p. 28.

Attention and Mental Representation

Earlier in this chapter we noted that in organizations, the task of observing job behavior or of acquiring information about the behavior of the people to be evaluated is likely to be one of the many things the evaluator is doing at any point in time. For example, your supervisor might decide to make a concerted effort to observe you doing your work, but he or she almost always has a number of other tasks to worry about, and it is unlikely that evaluators are able to devote their full attention to performance evaluation much of the time. Thus, the first question that might be considered in thinking about the cognitive processes that underlie judgments about performance is how evaluators sort the wheat from the chaff—what do they pay attention to and how do they mentally represent what behavior they have observed or what information they have acquired? Most generally, there appear to be two different cognitive processes for handling this task: automatic and controlled processes (Feldman, 1981; Kahneman, 2011). Familiar stimuli are usually processed automatically, with strong reliance on existing patterns and structures (e.g., categories and prototypes) while more novel or unexpected stimuli are often processed using a more controlled, conscious, and effortful process.

A category is a group of related objects that takes the form of a "fuzzy set" (Rosch, 1977; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). That is, categories do not have rigid boundaries, but rather represent potentially unstable groups of objects that are held together by similarity or "family resemblance." A prototype is an exemplar of that category. It is an image that summarizes the typical and distinguishing features of a category. Thus, if the categories are "dependable economy car" and "American sports car," a Subaru and a Corvette might serve as prototypes of their respective categories. Finally, schema (or schemata) refers to higher-level memory structures that contain verbal or propositional information (Feldman, 1986; Ilgen & Feldman, 1983). Schema can be used to represent the self and/or others in familiar situations (Ilgen & Feldman, 1983). One type of schema, referred to as a script (Abelson, 1976) refers to a mentally stored representation of a familiar event (e.g., visiting a restaurant). All of these terms—categories, schemas, and scripts—are useful for understanding the way raters acquire and make sense of information about the performance of the individuals he or she is called upon to evaluate. In particular, the behaviors a rater has observed are not stored as simple mental movies that can be played back at some later time. Rather, the mental representation of what one has observed is strongly influenced by the categories and schema that are relevant or activated at the time of observation, and this initial representation will have a decisive effect on what is later retrieved from memory. For example, if your first impression of a new employee is that he or she is uncomfortable collaborating with peers, you will tend to pay more attention to and more easily remember behaviors that seem in line with this impression.

The most important theme of cognitive research on performance evaluation is that the process of information acquisition is active rather than passive. That is, the rater does not simply bring in all of the available information from whatever he or she has an opportunity to observe, but rather selectively attends to some features of the ratees and their behavior, and devotes little attention to others. If we focus for the moment on the perception of behaviors, cognitive research suggests that the attention we devote to any particular behavior is a function of three variables: (a) the behavior itself, (b) the context of observation, and (c) the purpose of observation. That is, some behaviors are likely to attract more attention, regardless of the context or purpose of observation.

Organizational norms and standards (see Chapter 7) define some behaviors as important, desirable, unacceptable, or forbidden. It also seems likely that behaviors that carry strong evaluative implications will receive attention, almost regardless of the local context (Murphy, 1982b; Wegner & Vallecher, 1977). That is, categorization is very likely to involve placing each person along a good–bad, preferred–non-preferred continuum. Cognitive models have generally assumed that evaluation was the end product of the cognitive process, but there is evidence that evaluation is primitive and universal, and that evaluation may occur at the same time as, or even precede, other information-processing activities (Kim & Rosenberg, 1980; Osgood, 1962; Wegner & Vallecher, 1977; Zajonc, 1980). Research by Hastie and Park (1986) suggests that evaluations are stored at the time that behavior is observed, and that subsequent memory may be for evaluations rather than for behaviors.

Research on the effects of context on the cognitive processes involved in performance judgment suggests two conclusions. First, the salience of most behaviors (i.e., the likelihood that they will be the focus of attention) varies across situations (McArthur, 1980; Taylor & Fiske, 1978). In part, this is probably due to differences in the evaluative implications of specific behaviors in different situations. For example, a loud, verbally aggressive style of conversation may not attract any attention among a group of heavy equipment salesmen, but might be very noticeable in a receptionist; the behavior is appropriate in one situation, but not in others. Second, distinctive, novel features of the ratee or of his or her behavior will be highly salient (Langer, Taylor, Fiske, & Chantowitz, 1976). Behaviors that are infrequent might become salient through their novelty; behaviors that are important but commonplace may attract less attention.

Research on the purpose of observation, sometimes referred to under the heading of observational goals, is very relevant for understanding what behaviors will or will not be attended to by raters in organizations. Supervisors typically face multiple task demands and rarely have the luxury to devote all of their attention to behavior observation and evaluation (Balzer, 1986; Murphy, Philbin, & Adams, 1989). Thus, raters in organizations are likely to acquire information about ratees' performance while they are concentrating on tasks other than evaluation.

Even when raters do consciously seek out information about ratees, the purpose of observation will affect their information-acquisition activities. In general, raters will seek different information if they want to find out about people, tasks, or characteristics of the group (Murphy, Garcia, Kerkar, Martin, & Balzer, 1982; Williams, DeNisi, Blencoe, & Cafferty, 1985). The information they attend to will depend largely on what goals they have and on what else they are doing when they observe the behavior of the individuals they are asked to evaluate (DeNisi, Cafferty, & Meglino, 1984; DeNisi & Peters, 1996; DeNisi, Robbins, & Cafferty, 1989). For example, Kinicki, Hom, Trost, and Wade (1995) suggest that variations in the rating task and the rating environment might prime different categories (i.e., make them easier for the rater to access and use at the time when performance is being observed), and that the information that is most likely to be recalled at some later point is information that is related to and consistent with the categories that are most accessible to the rater while he or she is observing performance. Hauenstein (1992) suggests that subjects' expectations about the rating task (e.g., whether they will provide face-to-face feedback to ratees) affect encoding of performance information, and subsequently affect ratings.

Memory: Storage and Retrieval

A dominant theme of much of the research on memory in the last 30–40 years has been the fallibility of human memory. This should not blind us to the fact that people are capable of remembering a great deal of information and of retrieving detailed and accurate information that was stored in memory a long time ago. Unfortunately, people have a notoriously difficult time distinguishing between accurate memories and memories that *seem* accurate, in part because they can be so vivid and detailed. Peoples' belief that they can accurately recall what they have seen is an important explanation for the fact that so many managers and so many organizations continue to rely on subjective evaluations of job performance. That is, most of us believe we can accurately observe and recall behavior and that we can therefore form reasonably accurate evaluations of the performance of those we are asked to evaluate (Simons & Charbris, 2011), and therefore do not doubt our ability to fairly and accurately evaluate performance. Raters may lack confidence in *others'* ability to observe, recall, and evaluate the behavior of their subordinates, but they rarely doubt their own ability to execute this daunting task.

In Table 5.1, we noted that one of the steps in forming judgments about a ratee's performance is to retrieve information about his or her behavior from memory. One of the dominant themes of research on memory for the behavior of the people one is asked to evaluate is the decisive role of general evaluative impressions—that is, the global impression that a particular person is a good or a poor performer—in structuring what one recalls from memory. There is compelling evidence that memories of and evaluations of specific performance episodes are influenced by general evaluative impressions (Murphy, Balzer, Lockhart, & Eisenman, 1985; Murphy, Gannett, Herr, & Chen, 1986; Smither, Reilly, & Buda, 1988), and that we will more readily recall behaviors that are consistent with these general impressions than other behaviors. Indeed, there is evidence that raters may infer behavioral details from their evaluations, rather than (as is generally assumed) basing their evaluations on the total set of behaviors they have observed (Murphy, Martin, & Garcia, 1982). That is, if you believe that a particular subordinate is a generally good performer: (1) you will more readily recall instances of good performance than instances of poor performance, and (2) you may even believe that you remember positive behaviors, even if you never actually observed those behaviors. This reliance on general evaluative impressions in retrieving information from memory is one of the key reasons for the "halo effect"—that is, the tendency to rate specific aspects of job performance on the basis of overall judgments about whether the employee is a good or a poor performer rather than on the basis of information about the specific performance dimensions being rated. We will discuss halo errors in more detail in Chapter 11.

The ratings that are given to past performance are likely to influence ratings of future performance (Murphy, Balzer, Lockhart, & Eisenman, 1985; Reilly, Smither, Warech, & Reilly, 1998). First, previous performance influences the rater's general impressions about the ratee, and these impressions have a considerable effect on memory for behavior and on the evaluation of that behavior (Martell & Evans, 2005; Murphy & Balzer, 1986; Murphy, Martin, & Garcia, 1982; Woehr & Feldman, 1993). Slaughter and Greguras (2008) suggest that escalation of commitment might also represent a partial explanation for this finding. They note that once people are committed to a judgment or decision, they tend to increase their commitment over time, even in the face of clear evidence that the initial judgment or decision was wrong. So, if you start out thinking that Jennifer is a good performer, you might resist changing that evaluation, even in the face of behavioral evidence that suggests poorer performance. Over time, you will probably have a harder time recalling Jennifer's failures than her successes.

Similarly, the performance of other ratees can influence evaluations of the performance of a target ratee. There is evidence of both assimilation and contrast effects in performance ratings (Becker & Miller, 2002; Murphy, Gannett, Herr, & Chen, 1986; Sumer & Knight, 1996). Assimilation effects occur when your evaluation of a person is pulled in the direction of evaluation of others, whereas contrast effects occur when differences between the person to be rated and others are magnified. Assimilation is more likely when there are small differences between the people to be rated, whereas contrast effects are more likely when the differences are large enough to pay attention to.[2] On the whole, contrast effects appear more common and more difficult to eradicate (Jennings, Palmer, & Thomas, 2004; Maurer, Palmer, & Ashe, 1993; Palmer, Maurer, & Feldman, 2002; see, however, Becker & Villanova, 1995, who argue that many lab studies overestimate the strength of contrast effects).

As DeNisi (2006) notes, although general evaluations are critical determinants of what is or is not recalled, the situation is not always that simple. It is true that it is often easier to recall behavioral information that is consistent with general evaluative impressions, but sometimes the opposite can occur. That is, behavioral incidents that are *sharply* inconsistent with overall impressions might receive additional attention and might be more easily recalled. However, available research supports a general trend of memory biases in favor of behaviors that are consistent with your general impressions of a ratee's overall level of effectiveness.

A second major theme of research on memory for behaviors is that memory is dynamic rather than static. That is, memory is *not* like a bank vault, where items are stored for some period of time and later retrieved in their original form. Rather, memory can be influenced by task demands, and information that is stored in one way might be changed if information-processing objectives are changed. For example, Williams, DeNisi, Meglino, and Cafferty (1986) studied the way memory for behaviors changed if behavioral information was obtained for one purpose (e.g., identifying the best performer in a group) and subsequently used for another purpose (e.g., comparing the performance levels of all of the members in a group), and presented clear evidence that the new purpose could change the way behavioral information was stored in memory.

Information Integration

At some point, a rater who is asked to evaluate a particular subordinate's performance over the course of a year (or over the life of a project) will have to integrate a large body of information, some of which is likely to be based on direct observations of performance and other based on indirect sources (e.g., input from other supervisors, inferences based on the results of behavior). The set of available information is likely to include both instances of effective performance and instances of less effective performance, and for all but the best workers, it is likely that a thorough review of performance will include both positive and negative information. The question of how this information is put together to arrive at a judgment about the ratee's performance is an important one that has been extensively studied.

There is a long history of research in judgment and decision making showing that negative information tends to get disproportionate weight when decisions are made; DeNisi, Cafferty, and Meglino (1984) review this research, with special attention to its implications for performance appraisal. Ganzach (1995) analyzed several data sets and presented consistent evidence that when performance profiles include both positive and negative elements, the negatives receive more weight. The one exception is for performance profiles of high performers. These individuals may have their relative strengths and weaknesses, but their weaknesses are not instances of poor performance (i.e., negative information), but rather information that is positive but not as stellar as information about strengths. When all of the information in a performance profile is generally positive, there does not appear to be any tendency to over-weight the least positive dimensions.

One of the challenges in forming judgments about overall performance levels is that performance is not constant; all workers have good days and bad ones (Zyphur, Chaturvedi, & Arvey, 2008). Raters who are asked to evaluate their subordinates are faced with the difficult task of pulling together information about performance that is sometimes better and sometimes worse, and the different patterns of performance people show can pose particular problems for raters (Reb & Gregarus, 2010). For example, workers whose performance is improving probably benefit from this upward trend, and will be evaluated more positively than other workers whose average level of performance is just as good (Lee & Dalal, 2011).

The way performance information is conveyed to supervisors can influence their evaluations. Consider the two reports shown in Table 5.2. In the success report, it is clear that both John and Sam performed well, and they might both be regarded as highly successful and pretty similar in their performance levels. In the failure reports, a supervisor might correctly infer that John is twice as likely to fail as Sam and (ignoring the high success rate each achieves) conclude that there are large differences in their performance. Drawing on Prospect Theory (Kahneman & Tversky, 1979), Wong and Kwong (2005) presented evidence that differences in the likelihood of failure get more attention than corresponding differences in the likelihood of success when evaluating performance (See also Lee & Dalal, 2011).[3]

Table 5.2 Performance Reports That Emphasize Successes Versus Failures

Success	John was successful in 90% of the tasks he attempted. Sam was even more successful, achieving success in 95% of the tests he attempted.
Failure	John failed in 10% of the tasks he attempted, whereas Sam failed in only 5%.

Liking and Emotion: Affective Influences on Performance Appraisal

It is clear that there is a strong relationship between liking and performance ratings (Judge & Ferris, 1993); a meta-analysis by Sutton, Baldwin, Wood, and Hoffman (2013) suggests a very strong correlation between the rater’s liking for particular ratees and the performance ratings they receive (liking accounts for over 60% of the variance in overall performance ratings). This correlation might be interpreted as evidence of bias, but only if liking has nothing to do with performance and effectiveness on the job. Sutton et al. (2013) review several studies that suggest that the correlation between liking and performance ratings is not simply due to bias. First, a supervisor’s liking for particular subordinates is quite likely to be partially driven by their performance and effectiveness. High-performing employees are vitally important to the organization and they directly help supervisors accomplish the tasks and goals assigned to their work group. Second, raters who like and trust particular employees are likely to give them information, support, and assistance that allows them to perform well (Graen, 1976; Graen & Uhl-Bien, 1995; Kacmar, Witt, Zivnuska, & Gully, 2003).

In the previous section of this chapter, we discussed cognitive processes that influence the judgments raters make about the performance of the individuals they are called upon to rate. Interpersonal affect (liking versus disliking ratees) also influences judgments about job performance (Varma, DeNisi, & Peters, 1996), but the precise mechanism by which this occurs is not clear. One possibility is that there is a strong affective component to overall impressions of ratees (Zajonc, 1980), and that liking a particular ratee leads to the inference that he or she is a good performer, much in the same way he or she is seen as having a number of other positively valued traits. Lefkowitz (2000) reviewed research showing the affective regard is related to higher appraisals, but also to higher interdimensional correlations. The effect of liking on these correlations is consistent with the overall impression hypothesis.

Another possibility is that raters pay more attention to behaviors that are consistent with their general affective response, and therefore will pay attention to positive performance if they like one ratee, and attend to negative performance if they dislike another. Still another possibility is that raters rely on different prototypes and mental schema when thinking about ratees they like than when thinking about ratees they dislike, and that they are biased in favor of recalling behaviors that are consistent with these prototypes (Kinicki, Hom, & Trost, 1995; Robbins & DeNisi, 1993, 1994).

Robbins and DeNisi (1994) suggest that affect has a direct effect on performance evaluations, in addition to its effects on cognitive processes. Interpersonal affect is likely to be a performance cue, albeit an indirect one. The rationale here is that subordinates who are consistently ineffective, and therefore fail to make an adequate contribution to the success of a work group will, in the end, be disliked. Although interpersonal affect is not solely instrumental, and it is entirely possible that an ineffective worker will still be liked, it is reasonable to assume some link between effectiveness and liking. Interpersonal affect also influences both attention and memory, and it is likely that raters who like a particular ratee will find it easier to bring to mind instances of good performance.

Finally, affect is likely to play a role in bias in performance appraisal. In Chapter 11, we will review research suggesting that performance appraisals are not strongly influenced by the age, gender, or race of the ratee, but in cases where there is animosity or dislike of a ratee, the probability that demographic differences between raters and ratees will lead to lower ratings probably increases (Fiske, Harris, Lee, & Russell, 2016). It appears likely that affect can influence discrimination against members of various demographic groups, and liking an individual may reduce the effects of stereotypes, while disliking may make reliance on negative stereotypes more likely.

Finally, we should remember that “affect” is a fairly broad term. Performance appraisal research that has examined affect has generally concentrated on liking (Cardy & Dobbins, 1986; Dobbins

& Russell, 1986). Liking can be thought of as directed affect—it represents an emotional reaction to a specific person. However, not all affect is directed toward a specific person. Rather, there are at least two categories of undirected affect: (a) mood—which represents transient undirected affect, and (b) temperament—which represents chronic undirected affect. It is likely that both mood and temperament influence evaluations. A rater who is in a good mood at the time of observation and/or evaluation may give more positive evaluations than one who is in a sour mood. Raters who have a very positive, upbeat disposition may evaluate performance differently than raters whose temperament is surly or mean.

There have been comparatively few studies on the way mood or temperament influence performance appraisals. Sinclair (1988) suggested that mood influenced the way behavioral information was encoded and stored in memory; his study suggested that the most accurate ratings are obtained from raters who are in less positive moods. This finding is hardly a fluke; there is evidence that negative moods improve problem solving (Barth & Funke, 2010), and that a range of emotions can enhance performance on tasks that require cognitive activity (Blanchette & Richards, 2010; Seo & Barrett, 2007).

Standards for Evaluating Performance

Earlier in this chapter, we noted that different people often come to different conclusions about whether or not a particular employee has performed his or her job well. There are two possible explanations for this disagreement. First, they may have different information about the individual, either because they have observed different things or because they have received different inputs about that person's behavior and effectiveness. Second, they may interpret the same information differently because they apply different standards for defining good or poor performance.

We can think about the term "performance standards" in two ways. First, there are the *implicit* standards raters use to evaluate performance (i.e., their definition of good, average, poor performance). Second, there are *explicit* performance standards. The U.S. Office of Personnel Management (1998) defines a performance standard as "a management-approved expression of the performance threshold(s), requirement(s), or expectation(s) that must be met to be appraised at a particular level of performance."[4] Both types of standards are important; implicit standards play a critical role in the judgments raters make about performance whereas explicit standards play a critical role in the *organization's* evaluation of that same performance.

Personal values and beliefs are likely to influence implicit performance norms and standards. Most workers and supervisors have well-developed ideas of what constitutes a "fair day's work" (Zaleznik, Christensen, & Roethlisberger, 1958). These ideas reflect the individuals' opinions about the amount of effort and production that should be exchanged for the pay and benefits associated with the job. General beliefs about human nature are likely to affect both the standards themselves and their application (Wexley & Youtz, 1985). For example, popular theories of management describe two sets of assumptions about human behavior that are likely to affect the choice of a leadership style (McGregor, 1960). Managers who ascribe to Theory X believe that workers are inherently lazy and that they must be motivated by external rewards and should be closely supervised in order to achieve acceptable levels of performance. Managers who ascribe to Theory Y believe that workers are intrinsically motivated and that they will respond most readily to challenges and opportunities for growth at work. A Theory X manager is likely to have a detailed, strict set of standards and is likely to give very low evaluations to behaviors that deviate from those standards. A Theory Y manager is likely to have standards that allow for greater latitude in behavior and that are less punitive with regard to behaviors that deviate from those standards.

Several models of judgment and evaluation include the assumption that evaluative judgments are made with reference to standards that may vary from rater to rater (DeCotiis & Petit, 1978; Higgins & Stangor, 1988; Ilgen, 1983; Sherif & Sherif, 1969). West (1998) suggests that these standards are part of the rater's broader implicit theory of performance. Different supervisors and managers sometimes have very different ideas about what constitutes good or bad performacne or about *why* people perform well or poorly (Heslin, Latham & VandeWalle, 2005; Heslin & VandeWalle, 2011), and these implicit theories can lead to strong disagreements about performance even if there are few disagreements about the behaviors that were actually observed.

Explicit Standards

There are a variety of methods that might be used in setting explicit performance standards in a workplace. Work that involves relatively routine and repetitive activity (e.g., retrieving materials from a warehouse) is amenable to systematic work study, which can be used to develop *engineered labor standards*. Goomas and Ludwig (2009) describe the steps in creating this type of standard. First, "a task is subdivided into its elements, and each element (e.g., travel time) is given a discrete value (e.g., allocated number of minutes or seconds to travel from Point A to Point B). Values are arrived at based on activity sampling, group sampling, and time-series studies, as employees in these large industrial settings perform their job duties. The time needed for a qualified worker to carry out a task is then established" (p. 246). Performance is then assessed by comparing the standard time required to complete tasks with the actual time a worker takes to complete these tasks.

Kane and Freeman (1997) proposed a more general method for setting performance standards that is potentially applicable across a range of jobs. They note that for most job tasks, it might be possible to scale task performance from the lowest level that would be tolerated by an organization without taking administrative action (e.g., a level of performance where you would seriously consider firing the employee in question) to essentially perfect performance (or perhaps the highest level recorded), making it possible to scale performance (regardless of the task or job) on a 0–100 (i.e., worst to best possible) scale, and to express individual performance in terms of the percentage of maximum performance that was actually achieved. This, then, would allow organizations to create standards that were comparable across jobs, units, and the like.

The Kane and Freeman (1997) model is mainly concerned with setting standards for performance-based pay, but there can be a wider array of uses for performance appraisal than simply pay administration (see Chapter 8), and standards might be set for reasons that have very little to do with pay. For example, the performance standards that are created and enforced in an organization can say a good deal about the climate and culture of the organization. Some organizations exist in reward-rich environments (e.g., they have the resources and features necessary to attract a high-performing workforce) and might choose to set high standards as a way of defining their identity as a hard-charging firm. In other organizations, it might not be possible for workers to approach the maximum possible performance level, because of constraints (e.g., lack of materials, information, or support) or lack of a high-quality pool of applicants and incumbents, and they might be forced to set lower standards.

Bobko and Colella (1994) note that there are a number of aspects of performance standards (e.g., who sets standards, specificity, rationale, difficulty) that influence reactions to performance standards. Standards are an important part of the whole performance appraisal process; appraisal systems that include many desirable features (e.g., frequent feedback, good communication between raters and ratees) may still fail if performance standards are seen as unreasonable. Inappropriate standards can lead to work overload (Brown & Benson, 2005), which in turn can have negative impacts on the physical and mental health of employees, as well as long-term degradation in their performance.

Setting Explicit Standards

Standards are sometimes set through formal negotiation, particularly when unions or other bargaining units are involved. Saal and Knight (1988) note that union contracts often include clauses concerned with

- determining disciplinary procedures,
- scheduling of work and overtime,
- determining work methods, and
- determining production rates.

Each of these clauses defines a particular standard. For example, negotiations over disciplinary procedures in organization determine what types and levels of behavior can be a cause for official sanctions, what time frames apply in applying discipline (e.g., maximum time that can elapse between the infraction and the response), and what procedures will be followed to determine whether discipline is required or allowed for specific incidents. Negotiations regarding schedules and overtime might determine both actual working hours and the degree of latitude in determining whether schedules are met (e.g., how late an employee has to be to be classified as late). Negotiations over work methods might determine the standard procedures that are followed. Negotiations over production rates are likely to define standards for desired production levels, as well as standards for defining the amount of variance in production that will be acceptable.

It is more common for performance standards to either be imposed by managers and supervisors or to be negotiated between managers and their subordinates. Non-managerial employees are likely to have less of a role in setting performance standards, and supervsior–subordinate differences in the perceptions of reasonable standards can be a source of stress and conflict in organizations (Motowidlo & Peterson, 2008). One way to reduce the potential for conflict is to allow employees to have some input in defining performance standards, a common practice in performance appraisal methods that is based on assessing progress toward specific performance goals.

Setting Performance Goals

In earlier chapters, we mentioned management by objectives (MBO), which involves creating explicit standards through negotiation, albeit not in the formal, adversarial mode that is typical of negotiations between management and unions. MBO often involves a process in which the employee proposes a set of standards or goals that will define performance on his or her job, the supervisor reviews the goals and suggests revisions (if needed), and the two parties negotiate to reach a set of mutually agreeable goals (Szilagyi & Wallace, 1983). Although MBO is no longer a widespread method of management, the process of employees and their managers jointly setting performance goals and metrics for determining whether or not these goals are met has become quite common in perfomance appraisal.

Research on goal setting suggests that participating in determining goals is critical to success (Locke, Shaw, Saari, & Latham, 1981). That is, goals have a greater impact on performance when the employee helps to determine the goals than when goals are imposed from above. Because goals represent one type of external performance standard, the question of how goals are set might be a critical one for evaluating a goal-oriented performance appraisal system, at least with regard to the acceptability of the system to those who are being evaluated. Participation in the goal-setting process is likely to lead to stronger perceptions that the system is fair (Korsgaard & Roberson, 1995). However, the negotiation of goals, metrics, and standards opens the way for manipulation of the performance appraisal system.

The interests of the employees being evaluated versus the organization may not always be in sync when setting performance goals. It is to the employee's interest to make sure that performance goals are achievable and that whatever metrics are used to evaluate performance are under the employee's control. A more cynical, but probably realistic view is that it is in the employee's interest to make goals as simple to achieve as possible. The organization, on the other hand, has an interest in making performance goals more challenging. That does not mean that they benefit by making goals as difficult as possible; research on goal setting has clearly established that performance is maximized when goals are difficult but attainable, and are accepted as reasonable by employees (Erez & Kanfer, 1983; Locke & Latham, 1990). One of the challenges of management is to negotiate performance goals that serve the interests of both employees and the organization.

One of the key principles of effective performance management is that performance goals and objectives should be aligned with the goals, objectives, and strategy of the organization. As we have noted earlier, this alignment is probably easier for jobs and units that are closer to the core function of the organization and more challenging for jobs that are closer to the periphery. For example, if the primary strategy pursued by an organization involves cost containment, it makes sense for performance goals and objectives to emphasize efficiency in the use of organizational resources, minimization of waste, and the like. In an organization whose strategy emphasizes developing innovative products and services, there might be more emphasis on creative problem solving and on generating unique solutions.

Summary

We started this chapter by describing two key questions in designing a performance appraisal system: who should be involved in appraisal and how will they obtain the information needed to form sound judgments about the performane of the people being evaluated. Supervisory ratings are still the norm, but many organizations include self-ratings or ratings from other sources (e.g., peers). There are unique advantages and disadvantages with each plausible rating source, and there is no clear concensus that one source is always best. This might suggest that we should simply obtain information from multiple sources, but the answer is not so simple.

It is clear that ratings obtained from different sources do not often agree. Self-ratings are usually higher than ratings from other sources. Peers, supervisors, and other potential rating sources show disappointing levels of agreement. One implication is that in a multisource performance rating system, it will be common to find that different raters have quite different opinions about the performance of the people they are asked to evaluate, and these differences can undermine the success of multi-rater systems.

The "how" question represented a dominant theme in research on performance appraisal in the 1980s and 1990s, when studies of cognitive and affective processes in performance judgment were common. Interest in these basic processes has waned somewhat, in large part because of the realization that performance ratings do not always correspond with judgments about performance, which implies that a better understanding of the cognitive processes involved in forming judgments about performance might not tell us much about performance appraisals in organizations. Nevertheless, these judgments are an important starting point for understanding performance appraisal, and research on cognitive and affective processes in evaluating performance has made a definite contribution. Most generally, this research tells us that the task of observing performance, retrieving relevant information from memory at the time performance ratings are needed, and pulling together the information from various sources and various points in time to arrive at an overall evaluation is a complex one in which global impressions are at least as important as specific observations of or recollections of behaviors.

Finally, the judgments about performance involve some sort of comparsion between what a person has done and a set of performance standards. These standards might be implicit, representing a rater's opinion about what represents a "fair day's work," or they might be explicit performance standards and goals that are defined by management, negotiated with the union, or jointly arrived at by the supervisor and the employee. In a well-designed performance management system, these goals and objectives will be aligned with and will contribute to the execution of the organization's key strategies for success in a competitive environment.

Exercise: Writing Performance Standards

One way to get a sense of the range of issues you might need to consider when evaluating a subordinate's performance is to try out some of the tasks that are involved in performance appraisal. One common task is to develop and articulate standards that can be used to evaluate job performance.

Performance standards fulfill two functions. First, they tell employees what they are supposed to do or acomplish. Second, they tell employees how their performance will be evaluated. In Chapter 2, we described how the broad goals and strategies of organizations are used in performance management systems to drive the goals and standards used to direct and evaluate the performance of individual employees, and the goals and strategies of the organization and the work unit are an important source of information for establishing performance standards, but these higher-level goals do not always provide all of the information you will need. The checklist below is useful for writing performance standards; we suggest you use it to write standards for a job you are familiar with.

Performance standards should be measurable, realistic, and clear, and they should describe the behaviors or accomplishments that characterize sucessful performance, as well as providing clear guidance about performance levels that would fail to meet or that would substantially exceed expectations.[5] In describing steps for writing performance standards, we will assume that a job analysis has been carried out and that the appropriate performance dimensions have been identified. Performance standards describe what people need to do to perform adequately in each of the performance areas.

Performance Standards Checklist

1. What behaviors and outcomes does each performance dimension involve?
2. How many performance levels will these standards describe?
    a. Acceptable versus unacceptable performance—will performance be described solely in terms of whether or not specific expectations are met?
    b. Graded performance levels—will performance at several different levels be described?
3. Which metrics are relevant for evaluating these behaviors?
    a. Quality—do stakeholders care how well essential tasks were performed?
    b. Quantity—do stakeholders care how much work was performed?
    c. Timeliness—are deadlines or effectiveness in meeting schedules important to key stakeholders?
    d. Cost-effectiveness—is efficiency in using organizational resources to accomplish tasks important to stakeholders?
4. What specific measures should be applied?
    a. Objective indices—what should be counted, how, and how often?
    b. Judgments—if objective counts are not available or sufficient, who should make judgments about quality, quantity, timeliness, and so on?

To give an example of how such a checklist might be used to develop performance standards, consider the performance evaluation system that was used to evaluate faulty performance at one of the universities one of the authors worked at. Faculty performance was evaluated in terms of performance in teaching, research, and service. The performance standards for research required faculty to publish a certain amount each year. There were both quality and quantity standards, in the sense that publication in top-quality journals counted for more total points than publication in lower-tier journals, and there were formulas for establishing equivalence among a range of possibilities (e.g., books versus journals, first author versus second or third author). The metrics were mainly objective, but there was always some judgment applied in evaluating the quality of different outlets. This particular performance standard involved accomplishment rather than behavior, in the sense that it was the total number of publications and not the time or effort these involved that counted.

To apply the checklist yourself, think about a job that O*NET describes as having bright prospects for employment: bicycle repair. This job involves a mix of key tasks, including

- installing vehicle parts or accessories,
- adjusting vehicle components according to specifications,
- explaining technical product or service information to customers,
- assembling mechanical components or machine parts, and
- aligning equipment or machinery.

If you write performance standards for these tasks, you will find that some are likely to involve objective metrics and others to involve subjective measures. Depending on the nature of the bike shop there might be different levels of emphasis on quantity, quality, and timeliness metrics; a custom shop might emphasize high-quality work, whereas a shop that caters to a large drop-in customer base might emphasize timeliness and quantity. If you put yourself in the position of a manager in a bike shop and write some potential performance standards that might be used to evaluate bike repair specialists, this will give you a concrete understanding of the range of issues that need to be considered when creating standards of this type.

Notes

[1] The study of false memories can be traced back to Freud and his contemporaries.

[2] Sumer and Knight (1996) note that when raters are asked to rate previous performance, contrast effects are more likely than assimilation.

[3] Image theory (Beach, 1990) makes similar predictions, and these predictions have been confirmed (Pesta, Kass, & Dunegan, 2005).

[4] https://www.opm.gov/policy-data-oversight/performance-management/performance-management-cycle/planning/developing-performance-standards/

[5] https://www.opm.gov/policy-data-oversight/performance-management/performance-management-cycle/planning/developing-performance-standards/