Murphy Ch 6

6 RATING SCALES AND RATER TRAINING

Learning Objectives

- 6.1 Learn how behaviorally based rating scales are designed and used
- 6.2 Understand how memory aids such as behavioral diaries can help improve the accuracy of ratings, and why their effects are often limited
- 6.3 Learn a variety of ways of incorporating performance goals into performance appraisal
- 6.4 Understand the strengths and weaknesses of rater error training programs
- 6.5 Examine methods used to train raters to provide high-quality feedback

In one of his most stirring orations during World War II, Winston Churchill once told Americans "Give us the tools and we will finish the job." Psychologists and human resource specialists have devoted nearly 100 years to giving organizations the tools to effectively evaluate job performance. One way to characterize research on performance appraisal prior to 1980 is as a continual search for better tools—better rating scales, better methods of rater training, and better rating aids. As this chapter will show, these efforts have helped to improve performance *appraisal* in a number of ways, but it is fair to say that modifications of rating scales do not do much to improve the quality of performance *ratings*.

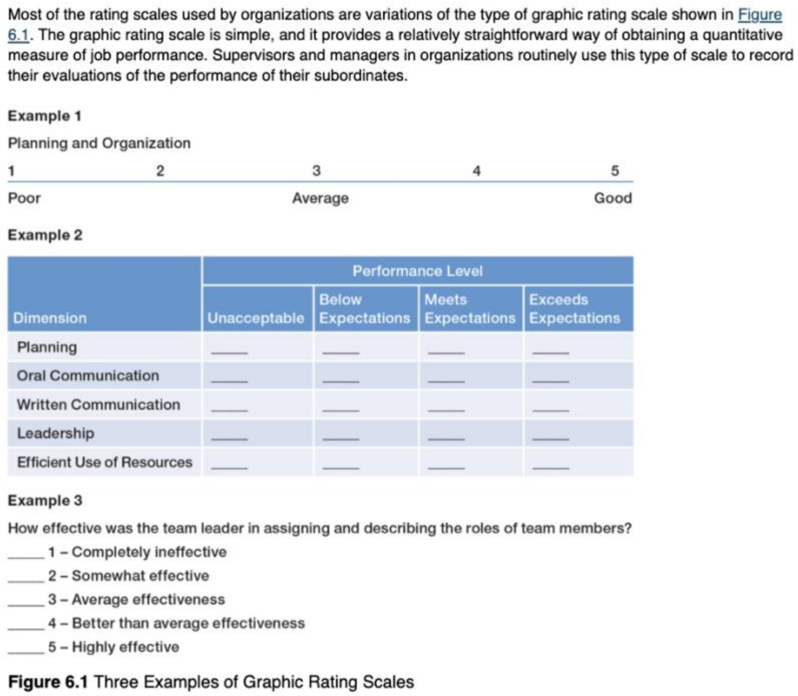
A second strategy for improving performance ratings is to improve the rater's capability to accurately evaluate performance. For example, a wide range of methods of rater training have been proposed, and there is evidence that training can help. We might also make the rater's task easier by introducing memory aids, such as behavior diaries. Again, there is evidence that these tools can help.

In this chapter, we start by reviewing research on efforts to develop better rating scales, particularly scales that include specific behavioral information. We also consider tools designed to assist raters in remembering the behavior of the employees they are asked to rate. We finish this section by considering alternatives to rating scales, and by considering how the incorporation of goals into performance rating changes rating processes and outcomes.

The second half of this chapter examines the impact of training on performance appraisal. We consider two very different strategies for training raters, one that focuses on minimizing errors and the other that focuses on maximizing consistency and agreement.

Rating Scales

Most of the rating scales used by organizations are variations of the type of graphic rating scale shown in Figure 6.1. The graphic rating scale is simple, and it provides a relatively straightforward way of obtaining a quantitative measure of job performance. Supervisors and managers in organizations routinely use this type of scale to record their evaluations of the performance of their subordinates.



Most of the rating scales used by organizations are variations of the type of graphic rating scale shown in Figure 6.1. The graphic rating scale is simple, and it provides a relatively straightforward way of obtaining a quantitative measure of job performance. Supervisors and managers in organizations routinely use this type of scale to record their evaluations of the performance of their subordinates.

Example 1

Planning and Organization

1 2 3 4 5

Poor Average Good

Example 2

	Performance Level			
Dimension	Unacceptable	Below Expectations	Meets Expectations	Exceeds Expectations
Planning	____	____	____	____
Oral Communication	____	____	____	____
Written Communication	____	____	____	____
Leadership	____	____	____	____
Efficient Use of Resources	____	____	____	____

Example 3

How effective was the team leader in assigning and describing the roles of team members?

____ 1 – Completely ineffective
____ 2 – Somewhat effective
____ 3 – Average effectiveness
____ 4 – Better than average effectiveness
____ 5 – Highly effective

Figure 6.1 Three Examples of Graphic Rating Scales

Figure 6.1 Three Examples of Graphic Rating Scales

The big advantage of a graphic scale is its simplicity, but this type of scale has many disadvantages. First, these scales do little to define the performance dimensions being rated. If different raters have different ideas about what “leadership” or “planning” mean, they are likely to disagree in their evaluations. Second, these scales do little to define what the performance levels being rated mean. Different raters might have very different ideas about what “average,” “good,” “meets expectations,” and others mean, and this could lead them to disagree in their evaluations. The ambiguity of both the performance dimensions and the performance levels included in these scales often make performance appraisals seem subjective and capricious. A substantial portion of performance appraisal research published during the 1970s and 1980s was concerned with attempts to develop better rating scales.

Before we discuss the different types of performance rating scales that were developed during this period, it is useful to think a bit about *why* scale development became such an important area of research. In our view, studies of rating scale formats were popular for a number of reasons. First, this is a topic that plays to the strengths of psychologists. That is, psychologists have a long history of involvement in scale development and measurement, and the development of rating scales for measuring job performance fit squarely within that tradition. Second, the development and refinement of rating scales provided what appeared to be a successful marriage of science and practice. It was easy for psychologists to develop, and for organizations to adopt new types of rating scales, and compared to many of the other interventions psychologists have developed, new rating scales seemed to be a simple application of scientific advances. Finally, the development of new and better rating scales seemed for a very long time to be a way psychologists could contribute to making the process of performance appraisal more fair and accurate and less fraught with conflict. Unfortunately, things did not turn out this way.

Behaviorally Anchored Rating Scales

In a highly influential article, Smith and Kendall (1963) described a procedure for developing clear, behavioral anchors for rating scales that could help reduce error and subjectivity in performance ratings, an approach they described as “retranslation of expectations.” This process involves input from several groups of potential raters. First, a group of supervisors, managers, or other subject matter experts is given a list of performance dimensions and asked to generate critical incidents—that is, examples of behaviors that indicate good or poor (or perhaps even average) performance on each dimension. Next, this list of behaviors is given to an independent group of supervisors, managers, or other subject matter experts who are asked to make judgments about which performance dimension each behavior example corresponds to. Next, this group, or perhaps another independent group of supervisors, managers, or other subject matter experts, is asked to indicate the performance level each behavioral example corresponds to. If these raters do not agree about which performance dimension or which performance level a behavior example corresponds to, that example is discarded. What is left is a list of behavioral examples that: (1) are written in language that makes sense to the supervisors or managers who will be using the scale, (2) provide clear examples of the dimension being rated, and (3) provide clear examples of specific performance levels. These behavioral examples are used as scale anchors; Figure 6.2 provides examples of behaviorally anchored scales developed by Murphy and Constans (1987) for evaluating the performance of lecturers.

In theory, behaviorally anchored rating scales (BARS) accomplish two distinct goals. First, they provide rating scales that help to define both what the performance dimensions mean and what the performance levels mean in clear behavioral terms. Second, they provide increases buy-in by involving a number of the potential users of the scales in the process of scale development. Even those raters and ratees who do not participate in scale development may view the scales favorably because of the heavy reliance of their colleagues’ feedback in developing scales.

The literature on BARS is voluminous (for a sampling of studies, see Bernardin, 1977; Bernardin, Alvares, & Cranny, 1976; Borman & Vallon, 1974; Hauenstein, Brown, & Sinclair, 2010; Jacobs, Kafry, & Zedeck, 1980; Schneider, 1977; Shapira & Shirom, 1980). However, this line of research eventually died out, for two reasons. First, as we discuss later in this section, by 1980 there was growing skepticism that adjustments to rating scale formats really made much of a difference in the quality of performance ratings. Second, there was evidence that BARS could introduce new biases in performance evaluation (Murphy & Constans, 1987; Piotrowski, Barnes-Farrell, & Esrig, 1989). Suppose, for example, that a store asks sales managers to rate their sales staff in terms of their success in maintaining good customer contact, and the rating scale uses the behavior “Calls customers to follow up, even after the end of the shift.” The sales manager observes one of her salespeople, who is usually lackadaisical in contacting customers, doing just this. There is evidence that observing this one behavior, because is corresponds with the behavior described on the rating scale, will bias ratings of customer contact.

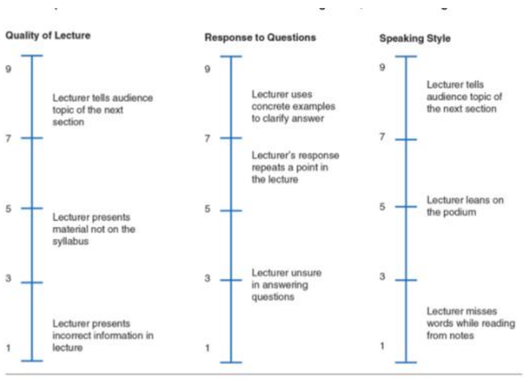

Quality of Lecture
9
Lecturer tells audience topic of the next section
7
5
Lecturer presents material not on the syllabus
3
Lecturer presents incorrect information in lecture
1
Response to Questions
9
Lecturer uses concrete examples to clarify answer
7
Lecturer's response repeats a point in the lecture
5
3
Lecturer unsure in answering questions
1
Speaking Style
9
Lecturer tells audience topic of the next section
7
5
Lecturer leans on the podium
3
Lecturer misses words while reading from notes
1

Figure 6.2 Behaviorally Anchored Rating Scales

Source: Adapted from Murphy & Constans (1987).

Figure 6.2 Behaviorally Anchored Rating Scales

Source: Adapted from Murphy & Constans (1987).

Behavior Observation Scales

A variation on the use of behavioral examples in evaluating performance is the behavior observation scale (BOS). A BOS presents a rater with a list of behaviors and asks how frequently he or she has observed each. Latham and Wexley (1977) present an example of a BOS item that might be used to evaluate the performance of a salesman—that is, “Knows the price of competitive products.” The rater would be asked to indicate how frequently this behavior has been observed, on the following scale: (1) Never (0–19% of the time), (2) Seldom (20–39% of the time), (3) Sometimes (40–59% of the time), (4) Usually (60–79% of the time), or (5) Always (80–100% of the time). They describe the development of BOS for logging supervisors that contained 78 behaviors measuring performance dimensions such as interactions with associates and crew, safety, and manpower and equipment management.

Proponents of BOS suggested that this method removes much of the subjectivity that is usually present in evaluative judgments. First, the scale was entirely made up of behaviors that are a normal part of the job. Second, the rater’s task was simplified, in the sense that rather than making an evaluative judgment about each behavior (e.g., Meets Expectations, Good, Bad), all the rater had to do was indicate the frequency with which each behavior was observed. Unfortunately, research into the cognitive processes involved in responding to BOS (Murphy & Constans, 1987; Murphy, Martin, & Garcia, 1982) suggests that the process of judging behavior frequency is every bit as subjective as the process of forming evaluative judgments. In fact, behavior frequency ratings may be more subjective than trait ratings or overall judgments; overall evaluations of the ratee’s performance appear to serve as a critical cue for estimating behavior frequencies. Thus, the use of BOS probably does not allow you to avoid the subjectivity of overall impressions or judgments.

Memory Aids

Behavior diaries have been suggested as a method of making performance appraisals more accurate and less susceptible to the effects of memory distortions (Balzer, 1986; Bernardin & Buckley, 1981). On the surface, the suggestion that raters should write down the behaviors they observe makes a good deal of sense, and there is evidence that these diaries can sometimes contribute to the accuracy of ratings (DeNisi & Peters, 1996). After all, it is difficult to remember behavior in any detail over long periods of time, and some of the inaccuracies that are introduced by relying on raters' memories might be reduced if good behavior diaries were available. Unfortunately, there are two problems with the suggestion that supervisors or managers should keep behavior diaries. First, the task of keeping detailed behavior diaries is time-consuming and burdensome. Suppose you supervised a dozen line workers and tried to write a diary entry every time one of them did something that was either effective or ineffective. You might not have time to do much else. Second, there is considerable evidence that the behaviors that *are* recorded in behavior diaries are not a representative sample of what the rater has observed. Even when simple checklists are used, there is evidence that overall impressions of ratees influence what is recorded or noted (Maurer, Palmer, & Ashe, 1993).

The suggestion that supervisors of managers should keep behavior diaries ultimately involves tradeoffs. These diaries take both time and effort, and the decision to collect this information can only be justified if the likely benefits outweigh this time and effort. There are two sorts of benefits that might reasonably be expected if raters kept behavior diaries. First, their memory for rate behavior *might* improve, which presumably would lead to better more accurate ratings. We are skeptical that the quality of ratings actually *will* improve. As we mentioned earlier, it appears that the same factors that bias memory for behaviors (i.e., a strong reliance on general impressions) are likely to bias behavior diaries, and if we encourage raters to keep these diaries, what we are likely to observe is a collection of behavioral incidents that are consistent with pre-existing opinions rather than an objective record of what the rater actually saw. A second benefit is perhaps more plausible. Suppose you maintain a detailed behavior diary. If, at the end of the year, an employee was not happy with his or her ratings (something that happens quite often), you could bring out the diary of behavioral records that support the ratings you gave. Of course, this feature of behavior diaries might only give the *appearance* of fairness if the record of behaviors you are able to produce is in fact a biased sampling of the real behavior of the employee. Nevertheless, the appearance of fairness is an important thing, and good behavior diaries could certainly enhance this appearance.

Evaluating Behavior-Based Rating Methods

Both BARS and BOS were designed to improve the quality of performance appraisals by making them more closely linked to the behavior of the individuals being rated and less subjective. Assessments of the effects of improvements in rating scales in general, and of the effects of using behavioral anchors in particular, suggest that these goals were not accomplished. There is compelling evidence that ratings are influenced by general impressions and global evaluations, and that by the time supervisors or managers get to the point that they have to evaluate performance over some period of time (e.g., an annual appraisal), the detailed behavioral information that might have once existed in the rater's memory is largely gone, and using behaviorally oriented rating scales is not enough to restore access to that behavioral information (Murphy & Cleveland, 1995).

More generally, the belief that improving rating scales would improve the quality of performance ratings simply did not pan out. DeNisi and Murphy (2017) reviewed nearly 100 years of research on performance appraisal and performance management. They noted that between 1917 and 1980, the most widely studied questions had to do with the hypothesis that some rating scale formats might be substantially superior to others as a means of collecting and recording evaluations of job performance. This hypothesis never received substantial support, and research on different types of rating scales eventually tailed off.

Landy and Farr's (1980) review was the first to reach the conclusion that adjustments to rating scale formats had few effects on the performance ratings individual employees received or on the quality of performance judgments. On the basis of the ongoing failure to demonstrate that changes to rating scale formats had much effect on the validity or usefulness of performance ratings they suggested a moratorium on rating scale research. This suggestion was highly influential (DeNisi & Murphy, 2017), and while there have been sporadic efforts over the years since 1980 to improve rating scales, this set of tools has received less attention in recent years.

It is useful to speculate on *why* over 60 years of scale format research failed to make a meaningful difference. It is likely that the answer is partly due to the way the cognitive processes involved in forming judgments about performance work. As we noted in Chapter 5, the processes involved in perceiving, categorizing, recalling, and integrating behavioral information are complex, and they are substantially influenced by overall impressions of whether a particular employee is a good worker or a poor worker. Behavior-based rating scales would make a lot more sense if performance judgments were built from the ground up—that is, if raters perceived and recalled behaviors pretty much as they happened and based their evaluation on the specific behaviors that were observed and recalled. It appears, however, that performance judgments are *not* simply an evaluation of the particular behaviors an employee has exhibited during the last year, but rather are influenced by conclusions you have already reached about that employee; the extent to which you like, trust, or depend on that employee; and the context within which evaluations occur.

More generally, the failure of rating scales to influence performance judgments is likely a symptom of a faulty diagnosis. That is, much of the research published between 1917 and 1990 was based on the assumption that performance ratings were not sufficiently reliable, valid, or fair because raters lacked the *ability* to form accurate judgments. Thus, it made lots of sense to try and develop tools to help the rater do his or her job. Since 1990, more emphasis has been placed on an alternative explanation, that raters lack the *motivation* to rate accurately (Banks & Murphy, 1985). Chapter 12 will examine this theme in more detail.

Scale formats might not have a large effect on reliability or validity of ratings, but behavior-based methods can have other effects, ranging from acceptance of ratings and satisfaction with performance appraisal to clarity regarding rating dimensions and levels (Tziner, Joanis, & Murphy, 2000; Tziner, Kopelman, & Livneh, 1993). There are three ways in which behavior-based rating scales might contribute to the success and acceptance of performance appraisal systems. First, as noted earlier, the process of developing these scales involves a good deal of input from many of the same people who will be using these scales in the future, something that is likely to enhance the perceived relevance and credibility of these scales. Second, they provide a common language for users to rely on when describing what performance dimensions and performance levels mean in a concrete sense. Finally, they *appear* to be more objective and fair than scales that provide no behavioral information at all.

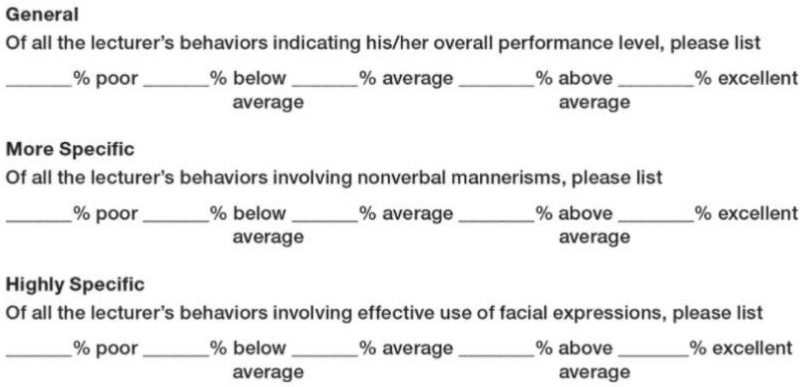
Performance Distribution Assessment

Job performance is rarely constant; any employee’s performance is likely to vary over time, and if the variability is extensive, a rating that captures the average level of performance may not be enough. A single number (the overall performance rating) can at best capture some aspect of performance (the average), while necessarily ignoring the rest. For example, suppose Sam shows average performance just about every day, whereas Frank shows excellent performance half of the time and terrible performance the other half. If they both received a single overall rating, they would receive the same score, but their performance is clearly not the same. To deal with this shortcoming, Kane (1983, 1986) proposed *performance distribution assessment* (PDA), in which raters are asked to describe the distribution of ratee performance rather than simply its average. In a more recent paper, Kane (1996) proposed a variety of metrics that might be used to characterize the distribution of job performance.

Distributional rating scales ask raters to indicate how frequently the individual being evaluated has demonstrated good, average, or poor performance. For example, Jako and Murphy (1990) describe performance distribution scales that might be used to measure teacher performance, and illustrate how this technique could be applied at several different levels of analysis, ranging from overall assessments to assessments of quite specific behaviors. These scales are illustrated in Figure 6.3.

PDA represents a more sophisticated version of the basic approach exemplified by BOS. In PDA, raters must indicate the frequency of different outcomes (e.g., behaviors, results) that indicate specific levels of performance on a given dimension. For example, the scale might describe the most effective outcome and the least effective outcome that could reasonably be expected in a particular job function, as well as several intermediate outcomes. The rater is asked to estimate the frequency of each outcome level for each ratee. One of the potential advantages of this format is that it allows you to consider the distribution or the variability of performance as well as the average level of performance in forming an evaluation. PDA involves some fairly complex scoring rules (a concise description of PDA is presented in Bernardin & Beatty, 1984; software now exists for PDA scoring), and results in measures of the relative effectiveness of performance, the consistency of performance, and the frequency with which especially positive or negative outcomes are observed.

Although the link is rarely made explicit, the use of distributional metrics rather than (or in addition to) the mean performance level to characterize job performance makes sense only if performance is dynamic. If most people perform at the same level most of the time, there would be no meaningful distribution of performance, and the mean would adequately capture the entire distribution. If performance is highly variable, the mean will capture only one aspect of the performance distribution.



General

Of all the lecturer's behaviors indicating his/her overall performance level, please list

______% poor ______% below average ______% average _______% above average _______% excellent

More Specific

Of all the lecturer's behaviors involving nonverbal mannerisms, please list

______% poor ______% below average ______% average _______% above average _______% excellent

Highly Specific

Of all the lecturer's behaviors involving effective use of facial expressions, please list

______% poor ______% below average ______% average _______% above average _______% excellent

Figure 6.3 Performance Distribution Scales at Three Levels of Specificity

Second, the type of distributional metrics that might be used depends substantially on the shape of the performance distribution. Suppose each person's performance varies from day to day or from week to week. If the distribution of performance is essentially normal, only two statistics, the mean and the standard deviation, will be needed to fully characterize the distribution of performance. If the distribution is skewed or irregular, more statistics might be needed, and if the distribution is sufficiently irregular, nothing short of a graph of performance levels might suffice.

The question of what might be accomplished using distributional performance assessments or what might be missed if organizations rely solely on the average level of performance probably depends substantially on the purpose of appraisal. If performance appraisals are used to set annual salary, the mean might be quite sufficient, but if they are used to provide feedback, it probably makes sense to give different performance feedback to Sam (who is always average) than to Fred (who is sometimes brilliant and sometimes awful).

Evaluating PDA

Evaluations of distributional rating methods have been somewhat mixed. Deadrick and Gardner (1997); Steiner, Rain, and Smalley (1993); and Woehr and Miller (1997) presented evidence supporting the validity and value of these methods, whereas Jako and Murphy (1990) found little evidence that these methods yield useful information. It is possible, as noted earlier, that the value of these methods depends substantially on the distribution of performance, and that the performance of specific individuals is more variable (or distributed differently) in some situations than in others. Woehr and Miller (1997) showed that distributional ratings often lead to similar conclusions to those reached using other rating methods about the average level of performance, but the additional information obtained by asking about rating distributions may help reduce the amount of measurement error in ratings.

Multi-Attribute Rating Systems

Manoharan, Muralidharan, and Deshmukh (2011) proposed an appraisal system based on multi-attribute decision making (MAUT).[1] They note that performance measures and judgments about performance are not always exact (e.g., a rater may conclude that Jeff is usually better at verbal communication than Dan), and they propose a complex application of fuzzy analytic tools to cope with this uncertainty. While their approach has the advantage of making the key issues in performance evaluation (e.g., identifying key performance dimensions, taking into account the intercorrelation among these dimensions, determining the weight to apply to each dimensions) explicit and mathematically rigorous, their approach would make performance appraisal more complex and less transparent. In particular, this approach requires complex calculations to translate the judgments of raters to actual performance scores, making it difficult for raters or ratees to understand precisely why an individual receives a good or a poor performance rating at any point in time.

The approach proposed by Manoharan et al. (2011) has not, to our knowledge, been implemented. While sophisticated in many ways, this approach is likely to cause more problems than it solves. In particular, this sort of system, in which the relationships between the evaluations of supervisors and the performance scores ratees actually receive are at best indirect is likely to make performance appraisal more confusing and less credible. We believe it is important to move in the opposite direction—to make performance appraisal transparent and easy to understand. Appraisal systems that involve setting and monitoring concrete performance goals are an example of this trend.

Using Performance Goals in Performance Appraisal

In Chapter 5, we noted that self-ratings are widely used in performance appraisal systems. The formats for these ratings vary from company to company, but these self-ratings often include two components: (1) articulation of performance goals and the metrics that will be used to assess the accomplishment of those goals, and (2) evaluations of the extent to which these goals have been accomplished. Performance goals are sometimes the result of discussions and negotiations between supervisors and subordinates. For example, an employee might be asked to propose performance goals and metrics for evaluating whether or not these goals are accomplished. These goals might be modified on the basis of input from his or her supervisor, or they might be accepted. This type of goal setting has some clear advantages. First, it directly involves the employee in determining what he or she will try to accomplish and how that accomplishment will be evaluated. Second, it provides a clear and objective basis for performance reviews. On the other hand, this goal-setting process provides a golden opportunity for the employee to "game the system." There is a clear motivation to try and set goals that are easy to meet, and even if your supervisor is well aware that you are trying to set easy goals, he or she may decide to play along; in Chapters 9 and 12 we will explore reasons why supervisors may choose to avoid conflict, even if it means lower performance levels from some employees.

In Chapter 2, we noted that the philosophy of performance management implies that goals will be imposed from above rather than being generated by the employees themselves. The reality is probably a bit more complex. That is, regardless of the performance management systems a company puts in place, it is probably common (and certainly smart) to get input from the employees and to take that input seriously. To be honest, despite the large literature on performance management, too little is known about how these systems actually operate. Nevertheless, a company that wants employees to actually accept performance goals will almost certainly want to give employees some genuine role in setting those goals.

It is widely recommended that performance goals and objectives should be written to be S.M.A.R.T.[2] That is, goals and performance objectives should be:

- Specific
- Measurable
- Agreed Upon and Achievable
- Realistic
- Time-Bound

Table 6.1 provides some examples of S.M.A.R.T. (and Not-So S.M.A.R.T.) performance objectives for a high school teacher. For example, the objective “All student exams will be graded and returned to the student within 5 days next semester” is specific (exams will be graded and returned), measurable (All … within 5 days), achievable and realistic (assuming that other demands of the job make this objective possible to reach and time-bound (next semester). In contrast, “Parents will become more engaged in their children’s education” is vague. It is not clear how engagement will be evaluated, or what the time frame is for its evaluation. It is not clear it is even a realistic goal; teachers do not typically have much control over parents’ level of engagement. Similarly, student scores on the state’s standardized graduation examination will increase by 10% next year may not be a S.M.A.R.T. objective. It is specific, measurable, and time-bound, but is it realistic and achievable? Depending on the test, it may or may not be possible to boost student achievement by this much.

Table 6.1 Examples of Performance Objectives for a High School Teacher

S.M.A.R.T. Objectives	All student exams will be graded and returned to the student within 5 days next semester Lesson plans will be drafted and submitted to senior teachers for comment at least one week before each class this year
Not-So S.M.A.R.T. Objectives	Parents will become more engaged in their children’s education Student scores on the state’s standardized graduation examination will increase by 10% next year

There is clear evidence that setting and monitoring performance goals can lead to increases in performance and effectiveness (Cunningham & Austin, 2007; Ivancevich, 1982; Locke & Latham, 1990; Locke, Shaw, Saari, & Latham, 1981), particularly when goals are accepted as legitimate and reasonable by employees (Erez & Kanfer, 1983). In addition to increasing performance, the process of setting and monitoring performance goals has the potential to increase the perceived fairness and accuracy of appraisals. These goals and objectives provide a clear idea about two things: (1) what is important to achieve, and (2) performance standards. If you receive a rating of “Meets Expectations” on “Oral Communication” from your supervisor, it may be hard to know precisely what this really means. On the other hand, a S.M.A.R.T.

objective not only tells you what it is you need to accomplish, it also provides a metric for evaluating whether you did in fact meet the goal.

One problem with goal-oriented appraisal systems is that they sometimes involve multiple, conflicting goals. For example, Cheng, Luckett, and Mahama (2007) examined the use of performance goals in measuring the effectiveness of call center employees. Performance measures included call time, net sales, adherence to schedule, accuracy in entering customer data into computer system, and call quality (e.g., is the caller polite and friendly?). They noted that these goals could come into conflict, especially when challenging goals were set. For example, if a goal of keeping call time as short as possible is set, this is likely to have an adverse effect on call quality. Another challenge of appraisal systems that incorporate specific performance goals is that they make failures more visible and less ambiguous. That is, if your performance plan includes a goal of increasing customer sales by 10%, and you do not reach that goal, it is clear to you and to your supervisor that you have failed. Unfortunately, failure to meet goals this year can have a negative influence on your performance next year. There is evidence that personality characteristics such as learning orientation can influence the way we react to failure and negative feedback, and that for some people, failure to meet a performance goal can lead to an avoidance of feedback and a propensity to minimize risk in the future (Cron, Slocum, VandeWalle, & Fu, 2005).

Another problem with appraisal systems built around the articulation and measurement of performance goals is one we have already encountered in our discussion of objective performance measures (Chapter 3)—that is, criterion deficiency. Appraisal systems built around objectives and metrics place a great deal of emphasis on those aspects of job performance that are most easily measurable, and place less emphasis on aspects of performance that are not so easily measured. In Chapter 3, we noted that job performance includes a good deal beyond simply accomplishing the tasks that are listed in your job description, and that important aspects of performance such as organizational citizenship or organizational deviance are not easy to translate into concrete, measurable performance objectives. Table 6.2 illustrates two S.M.A.R.T objectives that might be written to reflect important aspects of organizational citizenship and organizational deviance. While these objectives might be specific, measurable, achievable, realistic, and time bound, we doubt that anyone's performance objectives will include examples like these.

Finally, performance appraisal systems that incorporate S.M.A.R.T. objectives have the potential to change the power dynamics of performance appraisal, particularly if the goals and objectives are set by the individual employee. In particular, other than perhaps approving or editing goals proposed by the employee, a system built around S.M.A.R.T. objectives pretty much takes the supervisor or manager out of the picture in terms of performance evaluation and feedback. Suppose you have 10 S.M.A.R.T. objectives for the year. The whole point of S.M.A.R.T. objectives is to remove subjectivity about what it is people are supposed to do and whether or not they have successfully done it. Who needs a supervisor or manager to evaluate performance or to provide feedback? If the objectives are sufficiently wide-ranging to cover the full range of job tasks, the need for supervisory evaluations might be greatly diminished.

Table 6.2 Examples of Why It Is Hard to Write S.M.A.R.T. Objectives for Organizational Citizenship or Organizational Deviance

Citizenship	Will not complain or irritate coworkers during the next six months Will go above and beyond what is called for in job description at least four times during the next six months
Deviance	Will not steal materials or supplies worth more than $20 over the next six months Will not engage in bullying or sexual harassment during working hours for the next six months

Performance Appraisals Are Not Simply Numbers: Narrative Comments

Brutus (2010) notes that while most performance appraisals include both words (comments about performance) and numbers (numerical ratings of ranking), the narrative comments that are part of most appraisal rarely receive attention from researchers. She notes that narrative comments can add detail to numerical ratings (citing an example of an employee whose performance is below average, and who receives specific comments about performance deficiencies—in this case, an unwillingness to pay attention and act on input from coworkers), and that the recipients of feedback often pay more attention to the narrative comments than to the numerical ratings. She also notes that advances in research methods have made it increasingly practical to rigorously analyze the content of narrating comments and to use these analyses to help advance our understanding of performance evaluations. For example, she suggests that the amount and the specificity of commentary makes a difference (specific and detailed comments have potential value for identifying specific issues that represent strengths or weaknesses). Comments that provide suggestions are more useful than those that simply note problems, and comments that span the performance domain are more useful than comments that focus on a narrow slice of performance.

In their review of the performance appraisal system used by U.S. Navy officers, Bjerke, Cleveland, Morrison, and Wilson (1987) noted that raters sometimes use the narrative section to overcome the limitations of the numerical rating system. They noted that, like most organizations, performance ratings in the Navy tend to be inflated. If many officers receive high ratings, it can be difficult for raters to convey to their superiors that a particular officer really *is* excellent. Bjerke et al. (1987) noted that raters use narrative comments, offering faint praise to some officers ("looks good in his uniform") and strong and specific comments about salient performance dimensions for others ("his men will follow him anywhere"). Despite the clear importance of narrative comments, there has been little research on training raters to provide better or more useful narratives.

Research in performance evaluation has paid a great deal of attention, but there have been only a handful of empirical studies of the narrative comments that often accompany ratings (David, 2013; Gorman, Meriac, Roch, Ray, & Gamble, 2017; Wilson, 2010). This is a shame, because it is likely that employees pay more attention to the narrative comments they receive than to their numerical ratings (David, 2013). In some performance appraisal systems, narrative comments may be brief pro forma summaries of information already conveyed by ratings (e.g., a narrative comment of "Meets Expectations" for an employee who receives a rating in the middle of the rating scale), but in systems where comments are more extensive and individualized, narratives might provide critically important information to employees.

We see narrative comments as an untapped gold mine for advancing our understanding of performance appraisal. We suspect that numerical ratings have been more extensively studied precisely because they are numbers, which makes ratings relatively easy to study using a range of familiar statistical tools. Narrative comments are more difficult to analyze, but with the explosion of computerized text mining methods (Aggarwal & Zhai, 2012; Kuckartz, 2014), empirical research on narrative comments in performance appraisal is becoming increasingly feasible. So little is known about narrative comments, that it is difficult to do more than speculate about what an analysis of these comments might reveal, but we will offer two predictions. First, narrative comments might tell us a good deal about the rater. Some raters are probably more likely than others to give detailed, individualized comments. Second, they might tell us even more about the relationship between the rater and the ratee. We expect that high-quality narratives are most likely when the relationship between an individual rater and an individual ratee is either very good or (perhaps) very bad. High-quality supervisor–subordinate relationships should generate both more opportunities and a higher willingness to comment on the subordinate's performance. When the relationship is bad, raters might end up providing more detailed narrative feedback in an effort to either improve the relationship or to protect themselves by justifying poor ratings.

Ranking as an Alternative to Rating

In Chapters 1 and 3 we noted that one of the serious problems with performance ratings is inflation. It is not unusual for 80% of employees to be rated "above average," and the resulting tendency for most employees to receive pretty similar ratings severely undercuts many of the possible uses of performance ratings. One alternate is to do away with ratings altogether and ask supervisors or managers to rank their subordinates instead.

Ranking has been used in the performance appraisal systems in the military, usually as an adjunct to rather than a replacement for rating. Although ranking is generally regarded as a solution to rating inflation, even performance evaluation scales that rely on ranking rather than on rating are subject to manipulation. For example, during the 1980s, the performance appraisal form used by the U.S. Navy to evaluate the performance of its officers (Fitness Report) asked raters to provide performance ratings and also asked them to rank each ratee relative to his or her peers. So, if there were four lieutenants who reported to me, one of my jobs when completing the annual Fitness Report would to be rank-order these four.

On the surface, ranking should solve the problem of rating inflation. There is evidence, however, that raters "inflated" the rankings in the Fitness Reports they completed by using an unrealistic comparison standard, magnifying the number in the comparison group to make each actual subordinate look good (Kozlowski, Chao, & Morrison, 1998). For example, a ranking of third out of 15 looks a lot better than a ranking of third out of four, and by creating an inflated comparison group, a supervisor can avoid giving a low ranking to any of his or her actual subordinates.

The Fitness Report used by the Navy in the 1980s called for a full ranking of all subordinates. A more common application of ranking in performance appraisal is some sort of partial ranking, such as the forced distribution ranking systems that are used in many organizations (Grote, 2005; Guralnik, Rozmarin, & So, 2004; Pfeffer & Sutton, 2006; Stewart, Gruys, & Storm, 2010). Forced distribution ranking systems require supervisors to sort their subordinates into ordered categories, based on their overall performance or effectiveness. For example, a manager who has 10 direct reports might be asked to sort them into three categories, such as top 20%, middle 60%, and bottom 20%. That is, the supervisor might be asked to identify the top two performers and the bottom two performers; everyone else will end up in the middle category.

These systems provide a potential solution to some of the problems caused by rating inflation, by forcing raters to make distinctions between ratees when evaluating their performance. However, these systems are often disliked by raters (because they may not want to make these distinctions, especially if they believe that the differences between subordinates are small; Schleicher, Bull, & Green, 2009) and ratees alike. Blume, Baldwin, and Rubin (2009) studied reactions to forced distribution systems, and concluded that reactions are most likely to be negative when there are harsh consequences for ratees and when the group size is small. On the whole, rating systems (e.g., those that do not call for comparisons between employees) are seen as more fair than relative rating (ranking) systems, and forced distribution systems are seen as the least fair (Roch, Sternburgh, & Caputo, 2007).

Rank and Yank

General Electric, during the tenure of Jack Welch as its CEO (Welch & Byrne, 2001), instituted a forced distribution system in which a set proportion of employees (usually 10%) was forced into the lowest-performance category, and in which this lowest-performing group could be subject to a number of negative consequences, including dismissal. The “rank-and-yank” philosophy is based on the assumption that replacing the poorest-performing employees with new hires is an important part of maintaining and improving the effectiveness of organizations. While this system can seem cruel to the employees who are dismissed, proponents of rank and yank argue that keeping ineffective employees is ultimately harmful to the entire organization, and could even put everyone’s job at risk.

There is evidence that forced distribution rank-and-yank systems *can* help increase performance, especially if the identification of the best and worst performers is reliable and accurate (Scullen, Bergey, & Aiman-Smith, 2005). The effects of this method are especially dramatic in the first few years it is applied, but the payoff associated with these systems declines over time. Giumetti, Schroeder, and Switzer (2015) note that a rank-and-yank policy can have potentially negative consequences for members of minority groups. Even though the average performance ratings for white versus minority employees tend to be similar (Chapter 11 examines demographic effects on performance ratings), it is often the case the members of minority groups will be slightly more likely to receive lower rankings, and even small differences can have a large effect on who is fired and who is retained (Giumetti et al., 2015).

To understand why rank-and-yank systems have a short shelf life, consider the scenario illustrated in Table 6.3. Suppose there are 100 employees in an organization, and their absolute performance level can be scored on a scale from 1–5, with an average score of "3." Suppose also that new employees this organization recruits will have performance scores of "3." In the first year of the rank-and-yank system, you dismiss 10 employees whose performance ranked lowest and replace them with new recruits. This will produce a boost in performance (the average performance score will go from 3.0 to 3.2), but this strategy will not work for long. The first year, you replace 10 people whose absolute performance levels is "1" with 10 new recruits whose performance level is "3." The next year, there are no longer any "1s," but you still benefit by replacing 10 of the "2" performers with 10 recruits whose performance level is "3." The year after that, you do the same thing, replacing 10 of the "2" performers with 10 recruits whose performance level is "3."

Table 6.3 How the Effects of Rank-and-Yank Systems Diminish Over Time

Performance Level	**Year 1**	**Year 2**	**Year 3**	**Year 4**
5	10	10	10	10
4	20	20	20	20
3	50	60	70	80
2	20	20	10	0
1	10	0	0	0
Average	3.0	3.2	3.3	3.4

Note: Each year, dismiss the 10 lowest-performing incumbents and replace them with new recruits whose performance level is "3."

By the time you get to Year 4 in this system, there are no longer any people in the organization who are truly poor performers (absolute performance scores of 1 or 2), and if you continue to rank and yank, you will get rid of 10 people who are "3" performers and replace them with 10 new "3" performers. You will not be any better off, and because you will need to spend time and money training new employees, you are probably even worse off.

Unsurprisingly, rank-and-yank systems are often unpopular with employees. After using variations on the rank-and-yank system for many years, even GE has decided to replace this system with one that emphasizes performance feedback rather than one that emphasizes identifying poor performers.[3] Blume, Baldwin, and Rubin (2009) note that reactions to forced distribution systems depend on a number of characteristics of these systems, including consequences for employees who are rated in the top and bottom categories and the size of the comparison group. Perspective also matters; managers are much more likely to be favorably inclined to forced distribution systems than are employees (Blume et al., 2009).

What Problems Does Ranking Solve and What Problems Does It Create?

Replacing performance rating with ranking has the potential to solve the problem of rating inflation, although as Kozlowski et al. (1998) note, the solution is far from foolproof. On the whole, however, we believe ranking creates more problems than it solves. First, ranking only makes sense if there are sharp and meaningful differences between individual employees (full ranking) or if there really are distinct classes of employees (e.g., top, middle, and bottom performers). Otherwise, ranking systems force supervisors or managers to make distinctions where there might be little meaningful difference between people. Suppose your three best subordinates are very similar, but the ranking system only allows you to designate two as top performers. Forced distribution systems often force supervisors or managers to group employees in ways that do not truly reflect their performance. Finally, forced distribution evaluation systems only make sense if you assume that all work groups have the same average level and the same distribution of performance. Suppose one manager really is better at his or her job and really succeeds in creating conditions that lead to high performance levels. That manager will be forced to give performance ratings that are indistinguishable from another manager who supervises a group of truly poor performers.

Ranking systems might be quite appropriate for some purposes. In Chapter 8, we differentiate between using appraisal to distinguish between employees (e.g., to determine who should get raises or promotions) versus distinguishing within employees (identifying individual strengths and weaknesses). Ranking is potentially compatible with between-employee uses of appraisal, but is likely to be of little value for providing feedback or for guiding development. Suppose you find out that, in your manager’s opinion, you are a better employee than Fred but not as good as Susan. This sort of information might not tell you much about how you might develop into a better performer.

We believe that framing this question in terms of rating versus ranking is the wrong way to look at things. A case can be made that if ranking is used at all, it should be used in conjunction with rating rather than as a replacement for performance ratings. Both rating and ranking methods have their strengths and weaknesses, and it is probably better to incorporate both techniques in performance appraisal than to rely on either rating or ranking alone.

Rater Training

All of the different approaches to improving rating scales, or to developing alternatives to rating described in the first half of this chapter, share the same general goal: making performance evaluation easier, fairer, more consistent, and more accurate. Behaviorally anchored rating scales were designed with the idea that the scale itself could make it easier for supervisors and managers to understand what different performance dimensions and performance levels meant. Rating systems that asked employees to articulate performance goals were designed in part to remove ambiguity in evaluation. Once goals and metrics were defined, there would be little doubt about whether or not they were accomplished. On the whole, these efforts to perfect rating scales and evaluation methods did make a contribution by increasing the transparency and apparent fairness of performance ratings, but it is clear that the effects of variations in rating scale formats on the consistency, accuracy, and validity of performance ratings is small at best (Landy & Farr, 1980).

A second strategy for making performance evaluation easier, fairer, more consistent, and more accurate involves rater training. This strategy is based on the assumption that raters may lack the knowledge or the skills to consistently and accurately evaluate their subordinates. A wide range of strategies for training raters has been developed, including new methods that use virtual environments as a tool for training managers about performance appraisal (Morse, 2010). Two particular strategies have been extensively researched, one aimed at training raters to avoid particular errors (such as giving high ratings to all of their subordinates), the other aimed at training raters to develop a common frame of reference when evaluating their subordinates.

Rater Error Training

It has long been known that a large majority of employees receive performance ratings well above the scale midpoint (this midpoint if often used to represent average performance), suggesting that raters are unduly lenient in evaluating their subordinates. It has also long been known that ratings of separate aspects of performance (e.g., planning, oral communication) tend to be highly correlated even though these are in theory quite different behaviors, suggesting that raters rely on general impressions (leading to what Bingham, 1939, first labeled halo errors).[4] One of the most widely studied methods of rater training is referred to as rater error training (RET).

In general, RET involves: (1) giving raters information about errors such as leniency or halo, (2) giving them feedback regarding the possible presence of these errors in the performance ratings they give, (3) helping raters diagnose these errors, and (4) encouraging raters to avoid these errors when evaluating their subordinates. Latham, Wexley, and Pursell (1975) and Pulakos (1984) describe RET programs that go beyond simply encouraging raters to avoid giving high ratings to everyone, or giving the same ratings on all aspects of performance; a careful RET program might involve multiple rounds of information sharing, practice, and feedback. Nevertheless, the goal of most RET programs is a negative one—that is, encouraging raters to avoid certain common pitfalls.

RET training is potentially vulnerable to what Wegner and Schneider (2003) describe as the "white bear problem." Suppose I tell you to try, for the next 10 minutes, to *not* think about a white bear. The harder you try, the more the white bear will dominate your thinking. Similarly, if I train you *not* to give high ratings to everyone or *not* to give ratings that are too highly intercorrelated, you may end up focusing more on what you are trying to avoid than on what you are trying to accomplish:, fairer, more consistent, and more accurate performance ratings.

There is evidence that if you train people to avoid giving high ratings or to avoid giving ratings that are too highly correlated, they will comply (Latham et al., 1975: Pulakos, 1984). That is, if I train supervisors not to give too many ratings of "4" or "5" and to give more ratings of "3," they will do pretty much what they are told (although this sort of training, like most other sorts of training, is likely to have effects that fade over time). However, many studies have shown that while RET can help reduce leniency or halo, it probably does not make performance ratings more accurate (Bernardin & Beatty, 1984; Bernardin & Pence, 1980; Borman, 1979; Hedge & Kavanagh, 1988; Landy & Farr, 1983; Pulakos, 1984; see, however, Stamoulis & Hauenstein, 1993).

Woehr and Huffcutt (1994) conducted a meta-analysis to explore the effects of different types of rater training on performance ratings. They found RET had no clear effect on the accuracy of performance ratings. However, they were able to distinguish between studies that conducted rater error training as it was originally intended (i.e., increasing awareness of common errors) and those that conducted inappropriate response training (i.e., identifying an alternative “correct” rating distribution). Training that focused on increasing awareness of rater errors had a large positive effect on rating accuracy (*d* = .76) whereas training that referenced alternative rating distributions led to a decrease in rating accuracy (*d* = –.20). These findings emphasize the importance of avoiding inappropriate response training and focusing instead on educating raters about common rating errors (Woehr & Huffcutt, 1994). This study also showed that raters who have received this training that was aimed at effectively distinguishing between various categories of performance (performance dimension training) were less susceptible to providing common ratings across all dimensions. Performance dimension training, however, was not significantly related to rating accuracy or leniency errors (Woehr & Huffcutt, 1994).

By the mid-1990s, RET was falling out of favor, in part because of the lack of evidence that it improved the accuracy of performance ratings. However, the larger concern was a philosophical one. RET is essentially aimed at convincing raters not to do things that seem to undercut the value of performance rating, such as giving everyone high ratings or giving essentially the same rating on every aspect of performance. Consensus has grown that a better approach is a positive one—a training approach that tells raters what they *should* do rather than warning them about what they *shouldn’t* do. Frame of reference training has emerged as a positive and effective alternative to RET.

Frame of Reference Training

The goal of frame of reference (FOR) training is to ensure that all raters use a consistent set of ideas about work performance to assess employees (Athey & McIntyre, 1987).

Spotlight 6.1 Justify Your Ratings

One approach to dealing with rating inflation has been to require raters to provide explicit justification for very high or very low ratings. For example, some rating systems require raters to provide a concrete rationale for extreme ratings (e.g., ratings of “1” or “5” on a 5-point scale). What happens if you impose this sort of requirement?

It appears that raters react in one of several ways when faced with this sort of requirement. Some do precisely what the system asks them to do—give concrete explanations for extreme ratings. Others appear to react with boilerplate—by giving explanations that are pro forma statements, such as justifying a rating of “5” (where “5” = far exceeds expectations) by using the statement, “This employee far exceeds expectations” when giving ratings of “5.” The most common reaction appears to be to simply avoid extreme ratings altogether, thus avoiding the need to give detailed justifications.

Convincing raters to avoid extreme ratings strikes us as a poor solution to a very real problem. One problem that many performance ratings face is rating inflation. Convincing raters to avoid giving the highest-possible rating *appears* to solve some problems (if raters avoid the top of the rating scale, the mean rating will indeed come down), but in fact it does very little good. If raters decide to avoid ratings of “1” or “5,” all they accomplish is to transform a 5-point rating scale into a 3-point scale, and it is a good bet that the mean rating will be very close to “4,” rather than to “5.” This “solution” merely kicks the can down the road, and it does little, if anything, to improve the ratings

The process begins with providing raters with common standards for evaluating performance, and training programs teach raters how to match specific behaviors to performance dimensions using these standards. These evaluations are then connected to numerical ratings on a rating scale (Bernardin & Buckley, 1981). During frame of reference training sessions, raters are often given the opportunity to discuss examples of behaviors that fit into each performance dimension and to practice using the common standards to make ratings for these behaviors. Trainers typically also provide feedback regarding how well the ratings reflect the intended frame of reference for performance (Woehr & Huffcutt, 1994). The goal of this method is similar to the goal of behaviorally anchored ratings scales—that is, to encourage raters to adopt a consistent understanding of performance dimensions and performance levels. That is, FOR attempts to put all raters “on the same page,” by making sure that they all have a consistent understanding of the work behaviors they are asked to evaluate. The assumption here is that if raters adopt a consistent and appropriate set of categories for organizing and communicating information about the performance of their subordinates, this should enhance accuracy in recalling, making sense of, and evaluating the performance of their subordinates (Athey & McIntyre, 1987; Day & Sulsky, 1995; Woehr, 1994).

As Gorman and Rentsch (2009) note, “the ultimate goal of frame of reference training is to train raters to share and use common conceptualizations of performance when providing their ratings” (p. 1336). Although there is considerable evidence that this method of training improves the accuracy in recognizing, recalling, and evaluating rate behavior (Arvey & Murphy, 1998; Roch, Woehr, Mishra, & Kieszczynska, 2012; Stamoulis & Hauenstein, 1993; Woehr & Huffcutt, 1994), the assumption that this method actually changes performance schema (i.e., conceptualizations of what represents good versus poor performance and what performance includes) has rarely been tested. Gorman and Rentsch (2009) presented the first direct evidence that the effects of FOR can indeed be explained, at least in part, by the fact that this method helps raters adopt a consistent view of what they are looking for and what represents good versus poor performance. Their study showed that: (1) FOR training leads raters to have more similar and more accurate schema regarding performance dimensions and levels, and (2) the increased accuracy of performance schema is directly related to increasing levels of rating accuracy.

FOR is designed to help raters develop a consistent mental model of the performance domain and of what good versus poor performance looks like. Uggerslev and Sulsky (2008) note that the effectiveness of this training depends in part on what the rater’s mental model looked like prior to training. On the whole, FOR is most effective when raters’ initial conceptualization is similar to the one taught by FOR. Raters who have a substantially different view of what good versus poor performance entails, or of what is included in the performance domain, gain less from FOR training.

One of the key components of FOR (i.e., helping raters develop common standards for evaluating performance) might also be relevant for reducing disagreements between supervisors and subordinates. Schrader and Steiner (1996) note that disagreements between supervisors and subordinates are not always driven by different perceptions of subordinate behavior; supervisors and subordinates often agree on *what* was done in the workplace. Rather, they might apply different standards when evaluating that behavior, and training programs that help them adopt common evaluative standards might lead to higher agreement between raters and ratees.

FOR training is a well-liked tool in organizations. Uggerslev, Sulsky, and Day (2003) suggested that positive reactions to this type of training might help to mitigate some of the negative connotations associated with performance appraisals that many employees and managers endorse. Further, Sulsky and Kline (2007) found that participants generally had favorable reactions after receiving frame of reference training. The results of their study also pointed to the idea that more positive reactions were associated with higher learning as a result of the training. This research is consistent with the emerging trend of widening the scope of rater training research to focus on contextual factors including ratee reactions.

Alternatives to RET and FOR

Sometimes raters have the skills and the information needed to accurately evaluate their employees, but they do not believe that they have these essentials. As a result, training programs have been developed to help increase raters' confidence that they *can* provide accurate evaluations. Hauenstein (1998) suggested that rater training programs should include content to increase raters' self-efficacy—that is, their ability to accurately evaluate and rate performance. This idea was originally proposed by Bernardin and Buckley (1981), but it was largely ignored in most training programs, which focused instead on the type of frame of reference training that was included in the same article. Neck, Stewart, and Manz (1995) discuss the concept of increasing rater self-confidence using what they described as self-leadership training. The idea behind this emerging theme is that raters often experience uncertainty about their ratings, leading them to be more susceptible to biases and other rating errors. Training to enhance confidence in ratings might therefore increase the accuracy of performance appraisal ratings (Hauenstein, 1998). Whether this training actually *does* increase the consistency or accuracy of ratings is difficult to determine, in part because of the difficulty in measuring the accuracy of ratings. (Chapter 11 examines measures of the consistency and accuracy of performance ratings.)

Martell and Evans (2005) introduced a new form of rater training referred to as source-monitoring training. The goal of source-monitoring training is to address biases that stem from rater expectations of employee behavior. In these training programs, participants are instructed during the rating process to only report on behaviors that induce vivid and detailed memories rather than focusing on behaviors that just seem familiar to the rater. The results of Martell and Evans's (2005) study suggested that raters are able to separate memories of behaviors that actually occurred from memories of behaviors that seem to be familiar but may not have actually occurred. Thus, source-monitoring training can be effective for separating accurate judgments from those influenced by raters' expectations of ratees' performance behaviors. Martell and Evans (2005) called for additional research to further explore the value of source-monitoring training, but the idea has received limited attention in the literature since then.

Coaching the Coaches: Training Raters to Provide Feedback

Performance appraisal involves two distinct steps. First, managers or supervisors much evaluate the performance of their subordinates and assign ratings, rankings, and/or narrative statements to each of the important performance dimensions. Second, they must give feedback to employees, often in the form of a face-to-face performance appraisal interview. There is a large and detailed research literature dealing with the first step (i.e., performance rating), which we have reviewed in some detail. There is much less empirical research on how to train raters to give useful feedback.

To be sure, there are excellent books and reports discussing both the process of giving and the process of receiving feedback (e.g., Gregory & Levy, 2015; London, 2003; Society for Human Resource Management, 2017; Stone & Heen, 2014), and practically every publication aimed at human resource professionals is filled with advice about giving feedback, but it is striking how little we know about *how* supervisors and managers are trained to give performance feedback to their employees. We know even less about whether this training works or what sort of training works best in different circumstances.

Much of the discussion of performance feedback in the practitioner literature falls under the heading of *coaching*—that is, giving employees information and help designed to allow them to grow and improve their performance. Coaching is recognized as a critically important management competency by the U.S. Office of Personnel Management (2017), which provides practical advice on the essentials of coaching. What is missing from most of the literature on coaching is a clear understanding of *how* supervisors and managers learn this essential skills or what steps organizations undertake to develop this skill.

It appears that many organizations treat skill in coaching and providing feedback as a form of tacit knowledge—a skill that is picked up and developed without any clear and explicit program of training. This strikes us as a mistake, and also as a void in the current literature. It would be wonderful to provide organizations with concrete, data-driven guidance on how to best develop skills in providing feedback, but the current literature just does not support any particular program of development. What we do know from the broader literature on skill development is that both practice and feedback are required to develop most skills. This does lead to one practical suggestion that we believe is amply supported by research: it is important to evaluate the quality of feedback given by supervisors and to give *them* feedback on the quantity and quality of the performance feedback they provide. Beyond this, the field is wide open, and we eagerly await the development of empirical studies of the methods organizations use (or fail to use) to train supervisors and managers in giving performance feedback.

Summary

Recognizing the shortcomings of many performance appraisals, researchers have attempted to increase the consistency, validity, and accuracy of performance ratings by giving raters better tools for evaluating their subordinates. Much of the performance appraisal research published prior to 1980 was devoted to one strategy for improving performance ratings that involved improvements to rating scales. Scales that included specific behavioral anchors (behaviorally anchored rating scales) or that asked questions about the frequency of particular behaviors (behavior observation scales) received a great deal of attention, but other variations in scale formats, ranging from performance distribution assessments to assessments that abandoned ratings in favor of full or partial ranking (i.e., forced distribution scales) methods also exist. Landy and Farr’s (1980) review of the literature suggested that the effects of improving rating scale formats were very small, and they called for a moratorium on future rating scale research. With some limited exceptions, this moratorium has held. It is recognized that adjustments to rating scales can sometimes have benefits (adding behavior anchors makes scales appear less subjective and increases their acceptance), but that improving the scale often does little to improve the quality of ratings.

Rater training provides a more successful avenue for improving performance ratings. Early attempts at reducing rater errors by instructing raters to avoid particular errors were at best partially successful. If you train raters not to give so many high ratings, they will play along, but there is little evidence that decreasing the number of high performance scores does much to increase the consistency or accuracy of performance ratings. Frame of reference training, which is designed to ensure that all raters share common conceptions of the meaning of the performance dimensions they are asked to rate and of the different levels included on rating scales, has been more successful. There is evidence that this method of training can increase accuracy in recalling ratee behaviors and evaluating that behavior.

Exercise: Developing Behavior-Based Rating Scales

One of the recurring problems in performance rating is that different managers, supervisors, or raters sometimes disagree about what particular performance dimensions (e.g., planning, communication with external customers) actually refer to, or about what constitutes good, average, or poor performance on these dimensions. Performance appraisal researchers have used several different strategies to attack this problem, notably training (e.g., frame of reference training) and scale development. On the whole, it does not seem that refinements in performance rating scales have had much impact of inter-rater agreement (Landy & Farr, 1980), but nevertheless, there are a number of reasons why it makes sense to develop behavior-based scales. Rating scales that include concrete behavioral examples may contribute to the perceived fairness and acceptability of performance ratings, and the fact that rating scale development often involves getting a number of users of the appraisal system to participate in scale development may contribute further to users’ willingness to participate actively in performance appraisal.

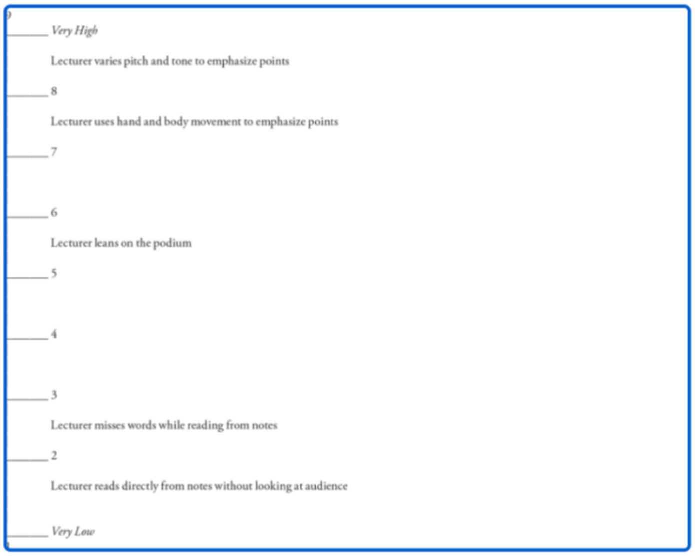
There is no set system for developing performance rating scales that include behavioral examples, but there are three general steps that are likely to be common to just about any process of behavior-based scale development. First, generate behavioral examples by asking participants in the appraisal process to provide examples of things that people do that illustrate different levels of performance or success in each of the performance areas that are rated. Next, use a “retranslation” process to identify the behaviors that are the clearest exemplars of particular performance dimensions and levels. For example, a group of users can engage in a linkage exercise, where they are asked to identify which performance dimension each behavioral example is linked to. If you have a reasonably large group of raters available, you can use a simple grid like the one below to record linkages.

Which dimension should each behavior be linked to? (check one)

Behavior	*Planning*	*Customer Communication*	*Time Management*
Sends e-mails that include incorrect information	______	______	______
Orders supplies before they run out	______	______	______
Is often late with reports	______	______	______

For example, if there are 10 people who complete this grid, and 7 or 8 link “Is often late with reports” to Time Management, that would constitute evidence that this is a good illustration of this particular aspect of performance. A second group of users can work with the list of behaviors that is most clearly linked to each performance dimension to determine what level of performance each example best illustrates. Here, you could use of similar grid, or you could ask each rater to place the behavioral examples somewhere on a 5-point rating scale, and then identify those ratings that can be most consistently scaled (e.g., items where the standard deviation of the performance level ratings is small). The result of these two steps will be a list of examples that clearly illustrate particular levels of performance on specific dimensions.

Murphy and Constans (1987) followed a similar process to create the scale shown in Appendix A, and illustrated below:



______ *Very High*

Lecturer varies pitch and tone to emphasize points

______ 8

Lecturer uses hand and body movement to emphasize points

______ 7

______ 6

Lecturer leans on the podium

______ 5

______ 4

______ 3

Lecturer misses words while reading from notes

______ 2

Lecturer reads directly from notes without looking at audience

______ *Very Low*

Notes

1. Murphy and Cleveland (1995) suggest MAUT as a framework for analyzing the outcomes of performance appraisal, but the Manoharad et al. (2011) paper suggests a broader embrace of the MAUT framework, using it as the basis for collecting performance evaluations.

2. Some practitioners have moved in the direction of advocating S.M.A.R.T.E.R goals, where E = Evaluate and R = Revise.

[3] http://blog.impraise.com/360-feedback/how-ge-renews-performance-management-from-stack-ranking-to-continuous-feedback-360-feedback.

[4] Halo and leniency errors are discussed in detail in Chapter 11.