11 EVALUATING PERFORMANCE RATINGS

Learning Objectives

- 11.1 Learn how inter-rater agreement measures are used to evaluate rating data
- 11.2 Understand why rater error measures were first proposed and why they turned out to be deficient measures of the quality of rating data
- 11.3 Learn how rating accuracy is defined and measured
- 11.4 Examine evidence regarding bias in performance ratings

Performance appraisal systems ask supervisors, peers, customers, and others to make judgments about the performance of the individuals or teams being evaluated; one question that immediately comes to mind is whether these judgments are correct, or at least, whether they are close enough to being correct that we should pay attention to them. This chapter examines the different methods that are used to determine whether performance ratings provide reliable, valid, and accurate information about the performance of the ratees who are being evaluated. Furthermore, we consider a question that is important but often overlooked in performance appraisal: whether performance ratings are useful to organizations and their members.

A simple example helps to illustrate the types of information that might be used to evaluate ratings and rating systems, as well as the range of questions that might be asked when evaluating the validity, accuracy, or usefulness of ratings. Suppose that there are two supervisors (Sam and Jose) who are asked to evaluate the performance of five new nurses on a hospital ward. Each nurse is evaluated on four performance dimensions: (1) quality of patient interactions, (2) timeliness in responding to calls and emergencies, (3) accuracy in recording patient information, and (4) making good decisions about treatments and medications. Each of these performance dimensions is rated on a 7-point scale (1 = falls far short of expectations, 4 = meets expectations, 7 = far exceeds expectations), and the overall performance score is a simple sum of these four scores. Table 11.1 shows the ratings Sam and Jose give. A quick look at the ratings in Table 11.1 tells you several things:

- Sam and Jose do not agree—Sam thinks Juanita is the best nurse and Pria is second worst, but Jose thinks Pria is the best and Juanita is second from the bottom.[1]
- Jose is a more lenient rater than Sam, usually giving higher scores.[2]
- Sam does not distinguish much among performance dimensions, giving pretty similar ratings to all four performance dimensions, but Jose is more likely to distinguish relative strengths from weaknesses.[3]

Table 11.1 Performance Ratings for Five Nurses

	Rater	
Nurse	**Sam**	**Jose**
April		
Quality	5	3
Timeliness	5	5
Accuracy	4	2
Judgment	4	4
Total	18	14
Pria		
Quality	4	7
Timeliness	4	6
Accuracy	5	6
Judgment	4	7
Total	17	26
Anne		
Quality	7	7
Timeliness	5	7
Accuracy	7	4
Judgment	5	6
Total	24	24
Jasmine		
Quality	3	6
Timeliness	4	5
Accuracy	5	7
Judgment	3	5
Total	15	23
Juanita		
Quality	6	5
Timeliness	5	7
Accuracy	7	4
Judgment	7	6
Total	25	22

All three of these observations (low agreement, differences in average ratings, differences in the variability of ratings) *might* be taken as evidence that the performance ratings in Table 11.1 are not good measures of job performance. After all, Sam and Jose are evaluating the same people, but they come to very different conclusions about their performance. These data do not

necessarily tell us, however, whether to believe Sam or to believe Jose (they disagree substantially), or even whether to believe either of them. The evaluation of performance appraisal systems, particularly of performance ratings or other measures of performance, is a complex undertaking, and there are a wide range of methods, indices, and metrics that have been used to evaluate ratings.

In this chapter, we will review four classes of measures that are used to evaluate rating data:

1. Measures of agreement and reliability
2. Rater error measures
3. Rating accuracy measures
4. Assessments of the validity and the usefulness of performance ratings

We will note that each of these classes of measures provides a partial answer to the questions of whether you should believe ratings like those shown in Table 11.1, and if so, whose judgment should you trust (Sam or Jose), but that in some cases, none of them fully answers these questions.

Do Raters Agree? The Reliability of Performance Ratings

Tests, measurements, and assessments of all types are known to be imperfect, and a substantial literature has developed describing ways of assessing and improving the quality of our measures. One of the most basic requirements of a measure is that it should produce consistent scores. So, if I evaluate your overall job performance today and next week, I should expect to obtain similar scores. If two different raters evaluate the same employee's performance, they should provide similar ratings. That is, tests, assessments, and measures such as performance ratings should be consistent, or *reliable*. Tests and measures that do not demonstrate this sort of consistency might be strongly tainted by measurement error, and measures that are mostly error will have little value to the individual or to organizations.

Measurement error is an important consideration in evaluating virtually all measures, including performance ratings. Viswesvaran, Ones, and Schmidt (1996) reviewed several methods of estimating the reliability (or the freedom from random measurement error) of job performance ratings and argued that *inter-rater correlations* provided the best estimate of the reliability of performance ratings. The basic idea here is that disagreements in the ratings that two separate raters give to the same employee are an indication of measurement error, and that inter-rater correlations therefore provide a means of estimating how much measurement error is present in ratings. In several subsequent papers, they elaborated upon the rationale for using inter-rater correlations as the best estimate of the reliability of performance ratings (e.g., Schmidt, Viswesvaran, & Ones, 2000; Ones, Viswesvaran, & Schmidt, 2008).

The correlations between ratings given to the same employees by two separate raters are typically somewhat low. Viswesvaran et al. (1996) suggest that the best estimate of the average inter-rater correlation for performance ratings is approximately .52. If these inter-rater correlations are used to estimate the reliability of performance ratings, one conclusion you would reach is that approximately half (48%) of the variability in performance ratings

represents random measurement error, while the other half (52%) is related to job performance.[4] A number of authors have argued that Viswesvaran et al. (1996) have greatly overestimated the role of random measurement error in performance rating, in large part because they have relied on an outdated approach to analyzing the quality of measurement (LeBreton, Scherer, & James, 2014; Murphy & DeShon, 2000). In particular, this approach to estimating the reliability of performance ratings is based on a theory of measurement that separates all of the variability in performance ratings into two categories: (1) true differences in performance, and (2) random measurement error. This approach is no longer accepted as reasonable by specialists in psychological measurement.

The methods proposed by Viswesvaran and colleagues for estimating the reliability of performance ratings are based on classical test theory (CTT), a theory that was once central to psychological measurement. The core idea of CTT is that observed scores (X) can be broken down into two independent components, true scores (T) and random measurement error (e), or X = T + e. This simple equation provided a starting point for a sophisticated theory of measurement and for many methods of estimating reliability (Lord & Novick, 1968). If we apply this framework to performance rating, this theory suggests that rater agreement can be explained in terms of true scores and that rater disagreement can be explained in terms of random measurement error.

By the 1970s, doubts about the usefulness of CTT were being raised (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). The main problem with this framework is that it ignores sources of variability in ratings that are not either a function of the thing you are trying to measure (true scores) or random measurement errors. For example, Murphy and DeShon (2000) noted that there are many reasons why raters agree that have nothing to do with job performance. For example, there is a well-known attractiveness bias (see, for example, Drogosz & Levy, 1996); attractive ratees tend to receive more favorable ratings, and this bias can lead to systematic agreement between raters, even if the raters know nothing about the performance of the people being evaluated. Similarly, there are many reasons that raters disagree, and some of these are systematic rather than being sources of random measurement error. Murphy and DeShon (2000) suggested that generalizability theory offers a more sensible model than CTT for analyzing the reliability of performance ratings.

Generalizability theory starts with an analysis of the different sources of variability in rating. For example, suppose two raters evaluated the performance of 10 subordinates (ratees) on four performance dimensions (e.g., Production Quality, Timeliness, Efficient Use of Resources, Communication). Table 11.2 lists the potential sources of variability in ratings.

Most important, generalizability theory does not require you to assume that the only reason raters agree is because of the true performance of the people being evaluated or that the only reason they disagree is because of random error. For instance, some raters give higher ratings than others. This might be evidence of disagreement if, as in the example illustrated in Table 11.1, they both evaluated the same ratees (in this example, Sam gave slightly higher ratings than Jose), or it might be evidence of systematic differences in the performance of the people being rated by different supervisors. The assessment of what represents true scores and what represents errors of measurement might depend on the way rating systems are designed (e.g., do multiple raters evaluate each ratee?) and used.

Table 11.2 Sources of Variability in Performance Ratings With Multiple Raters and Rating Contexts

Source	Meaning
Ratees	Some employees perform at a higher overall level than others.

Source	Meaning
Dimensions	The group of ratees might perform better on some aspects of the job (e.g., timeliness) than others.
Raters	Some raters are more lenient than others and give higher average ratings.
Rater X Dimension	Some raters will give high ratings on some dimensions and lower ratings on others.
Ratee X Dimension	Some ratees differ from others in terms of their relative strengths and weaknesses.
Rater X Ratee	Some raters will give high ratings to some ratees and lower ratings to others.
Unexplained Variability	Some of the variability in ratings has nothing to do with raters, ratees, or contexts, and probably represents random measurement that, taken together, resemble random measurement error.

In most organizations, the point of performance appraisal is to help evaluate who is doing well or poorly, and to evaluate peoples' strengths and weaknesses. If the purpose of rating is to help make between-person decisions (e.g., promotion, salary), variability due to ratees represents true score. In generalizability theory, variability due to other possible sources, such as differences between raters, differences between dimensions, or interactions (e.g., rater X ratee, and rater X dimension) are all potential sources of *systematic* measurement error. For example, in a multisource rating system, where each rater has a different perspective (e.g., supervisor, peer, subordinate), variability due to raters is expected and is not considered measurement error. If you expected supervisors, peers, subordinates, and others to all give similar ratings, there would be no real point in obtaining ratings from multiple sources (Ock, 2016).

Generalizability theory encourages you to carefully consider various explanations for the variability in performance ratings and to collect data that will allow you to estimate exactly how each of these effects will influence the generalizability of ratings. Generalizability depends substantially on exactly what you are trying to measure (e.g., between versus within-person differences), and you cannot simply assume that these effects are small or even that they are independent from one another or from true scores. Thus, generalizability theory requires a detailed and sophisticated understanding of how ratings will actually be used and interpreted before numerical estimates of reliability can be obtained.

A number of studies have examined the roles of systematic and random error in performance ratings, as well as methods of estimating systematic and random error (Fleenor, Fleenor, & Grossnickle, 1996; Greguras & Robie, 1998; Hoffman, Lance, Bynum, & Gentry, 2010; Hoffman & Woehr, 2009; Kasten & Nevo, 2008; Lance, 1994; Lance, Baranik, Lau, & Scharlau, 2009; Lance, Teachout, & Donnelly, 1992; Mount, Judge, Scullen, Sytsma, & Hezlett, 1998; Murphy, 2008a; O'Neill, McLarnon, & Carswell, 2015; Putka, Le, McCloy, & Diaz, 2008; Saal, Downey, & Lahey, 1980; Scullen, Mount, & Goff, 2000; Woehr, Sheehan, & Bennett, 2005). In general, these studies suggest: (1) that there are indeed systematic sources of measurement error that CTT overlooks, and (2) there is considerably less random measurement error in performance ratings than studies of inter-rater correlation would suggest. For example, Scullen et al. (2000) and Greguras and Robie (1998) examined sources of variability in ratings obtained from multiple raters. They found that about one-third of the variance in performance ratings obtained from multiple raters is likely due to ratee performance. There is both systematic and idiosyncratic rater variance present in ratings; about 15%

of the variance is likely due to differences in ratings due to perspective (i.e., differences in the ratings from raters at different levels in the organization). The largest source of variance in ratings is due to raters, some of which is likely due to biases or general rater tendencies (e.g., leniency).

Greguras, Robie, Schleicher, and Goff (2003) applied generalizability theory to the analysis of ratings collected from multisource rating systems that were used for either administrative or developmental purposes. Their analysis showed that a substantial portion of the variability in performance ratings in these systems is due to Rater and Rater x Ratee effects, and that variability due to performance differences across Ratees was smaller than variability due to Raters and Rater x Ratee interactions of unexplained variability in ratings (measurement error).

Dierdorff and Surface (2007) analyzed peer ratings collected in a variety of contexts, and while they did not apply generalizability theory in their analysis, their findings converge with those cited above, showing that there are strong and systematic context effects in peer ratings, and that these are distinct from both ratee performance and random error. Similarly, in their analyses of multirater systems, both Hoffman et al. (2010) and Woehr et al. (2005) report substantial source effects, indicating systematic differences in what ratings obtained from different sources (e.g., supervisors versus peers) measure.

In general, studies of the sources of variability in performance ratings lead to a few important conclusions:

- Differences between ratees are usually to be larger than differences within ratees. That is, performance ratings appear to more reliably tell us who is a better or worse performer than what strengths and weaknesses individual employees show.
- Differences between raters are, on the whole, at least as big as differences between ratees. This may be an effect of rating inflation; if most ratees receive high ratings, differences between ratees must be small.
- Ratings are not as unreliable as inter-rater correlations would suggest, but they often include multiple sources of variability that is probably not linked to the performance of the individuals being rated.

Rather than attaching a single reliability coefficient to ratings, O'Neill et al. (2015) suggest that we consider the inferences we are attempting to draw from ratings. Consider the following possibilities:

- Based on their overall average, can I conclude that Juanita is a better performer than Pria?
- Based on the overall averages, can I conclude that the overall performance of the four people who report to one supervisor is higher than the overall performance of the six people who report to another supervisor?
- Based on their scores on individual performance dimensions, can I conclude that Pria has different strengths and weaknesses than Juanita?

Different reliability coefficients would be used in answering each of the questions above, and generalizability theory provides a useful framework for calculating these coefficients.

Reliability of Multisource Ratings

Performance appraisal systems that use multiple raters, including but not limited to 360-degree rating systems, provide unique opportunities to isolate and reduce several potential sources of systematic and random measurement error in ratings. For example, in a traditional performance appraisal system, in which each employee is rated by his or her direct supervisor, it is hard to tell what ratings tell us about the rater versus the performance of the person being rated. If there are multiple raters, it might be possible to separate variability that is due to the rater from variability that is due to the person being rated, and in a multilevel system, it might be possible to go further, separating systematic differences in what is observed or the perspectives taken by peers, supervisors, subordinates, or customers from variability in ratings that is due to simple disagreements between similarly situated raters. Conway (1998) proposed and demonstrated methods for analyzing multi-rater or multilevel data and showed how these methods could be used to test specific hypotheses about biases in rating. Lance (1994) described a related model that could be used to identify rater and rating source biases in performance ratings.

There are two questions that are important to ask when comparing ratings from different sources: (1) are there systematic differences? and (2) do they agree? The answer is yes and no. There are systematic differences; self-ratings are generally higher than ratings from others (Valle & Bozeman, 2002). Agreement between subordinates, peers, and supervisors is typically modest, with uncorrected correlations in the .20s and .30s (Conway & Huffcutt, 1997; Valle & Bozeman, 2002). However, given the low levels of reliability for each source, it is likely that the level of agreement among sources is actually somewhat higher. Harris and Schaubroeck (1988) report corrected correlations between sources in the mid .30s to low .60s. Viswesvaran, Schmidt, and Ones (2002) apply a more aggressive set of corrections and suggest that in ratings of overall performance and some specific performance dimensions, peers and supervisors show quite high levels of agreement.

This disagreement between rating sources is not necessarily a bad thing. Different sources have very different information and perhaps different standards when evaluating performance; the rationale of multisource systems is that different sources *should* disagree and that by obtaining information from many sources, you will get the fullest possible picture of each ratee's performance (Borman, 1974; Bozeman, 1997; Gorman, Cunningham, Bergman, & Meriac, 2016; Hoffman et al., 2010). Nevertheless, these data do suggest that the common practice of relying on only one source (the direct supervisor) is unlikely to provide high-quality assessments of performance.[5] Hoffman et al. (2012) suggest that the quality of multisource ratings could be improved by adopting rating scales that more clearly define performance dimensions, but it is not clear that this will fully resolve the potential problems caused by differences in the perspectives of subordinates, supervisors, and peers.

Classical test theory suggests that adding more raters should increase both the reliability and validity of performance ratings, but as Howard (2016) has shown, adding more raters has limited effects on reliability and validity, in part because raters often show a mix of random and systematic disagreements. When you are developing an ability test, or a test of knowledge, it is sometimes possible to make the test more reliable by adding items that are essentially parallel measures (i.e., items that all measure pretty much the same thing). The same strategy is not

feasible in performance rating—raters are simply not parallel measures. Rather, each rater brings different information, perspectives, biases, and rating tendencies to the task, and these do not simply cancel out as you add more raters to the appraisal process.

Tornow (1993) asked the provocative question of whether multisource ratings should be thought of as a means or an end. That is, we could think of multisource rating systems as a means to get the most accurate assessment of a ratee's performance, in which case disagreements among raters could be a real problem. Alternately, we could think of multisource systems as a tool for development and discovery, in which disagreement among sources becomes just another data point that might prove useful. That is, it might be more useful to find out that your peers regard you as a poor communicator while your customers think you are quite good in communicating information and advice than to find out the average across all raters. Unfortunately, the data presented by Greguras, Robie et al. (2003) suggests that multisource ratings do not provide reliable measures of ratees' overall performance level. What should be a strength of a multisource system (obtaining data from multiple perspectives) is likely a weakness if these ratings are used to make distinctions between ratees.

It is commonly assumed that raters are more likely to agree on specific, observable aspects of behavior than on more abstract dimensions (Borman, 1979). Roch, Paquin, and Littlejohn (2009) conducted two studies to test this proposition, and their results suggest that the opposite is true. Inter-rater agreement is actually higher for dimensions that are less observable or that are judged to be more difficult to rate. Roch et al. (2009) speculate that this seemingly paradoxical finding may reflect the fact that when there is less concrete behavioral information available, raters might fall back on their general impressions of ratees when rating specific performance dimensions.

Other studies (e.g., Sanchez & De La Torre, 1996) have reported that accuracy in observing behavior is positively correlated with accuracy in evaluating performance. That is, raters who have an accurate recall of what they have observed do appear to be more accurate in evaluating ratees. Unfortunately, accuracy in behavioral observation does not appear to be related in any simple way to the degree to which the behavior in question is observable or easy to rate.

Rater Error Measures

There are several well-known patterns in performance ratings that suggest that these ratings are flawed measures. First, it is not unusual to find that 80–90% of all employees are rated as "Above Average." Second, differences in the ratings received by different employees often seem small, given obvious differences in their performance. Third, ratings on conceptually distinct aspects of performance are often highly correlated. One explanation for these phenomena is that raters make systematic errors in their evaluations of performance.

Discussion of the three most common rater errors—leniency, range restriction, and halo error—can be traced back over 60 years (Bingham, 1939; Kingsbury, 1922, 1933). These concepts still influence the ways in which rating data are analyzed (Saal et al., 1980; Sulsky & Balzer, 1988) and have provided the foundation for a large body of research. We find it useful to deal

separately with two classes of so-called “rater errors”: (a) distributional errors, and (b) correlational errors.

Distributional Errors

Examining the distribution of the performance ratings a particular rater assigns has long been thought to provide important clues to particular rating tendencies that might create problems in organizations and that might indicate symptomatic errors on the rater’s part. Suppose, for example, that there are three supervisors (Frank, Janice, and Al) in different parts of a manufacturing plant who each rate 10 employees, and the distributions of their ratings take the form shown in Figure 11.1.

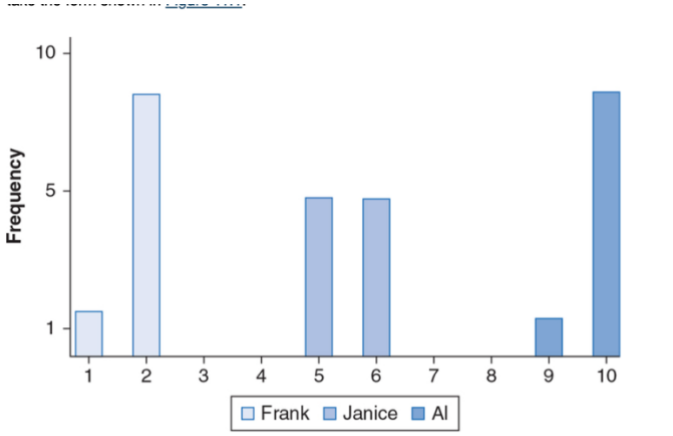

Figure 11.1 Rating Distributions That Illustrate Leniency, Severity, and Range Restriction

Figure 11.1 Rating Distributions That Illustrate Leniency, Severity, and Range Restriction

Frank gave low ratings to just about everyone. Janice gave ratings near the middle of the scale to everyone. Al gave high ratings to just about everyone. It is, of course, possible that all of Al’s subordinates are great workers, all of Frank’s are terrible, and all of Janice’s are truly average employees, but it is more likely that the ratings tell us more about the raters than about the employees being rated. Figure 11.1 suggests that Al is a lenient rater, while Frank is a severe one. It also suggests that Janice is unwilling to identify either high or low performance, and that none of the three raters makes distinctions among the 10 employees he or she supervises. That is, the three raters appear to be committing common rater errors. We say these raters *appear to* commit rater errors because we assume that there are usually some meaningful differences in the level of performance individual employees exhibit and that not everyone is either a wonderful or a terrible employee, but we do not necessarily *know* this.

The assumptions that underlie most measures of leniency, range restriction, and central tendency have been criticized. First, the true distribution of the performance of the group of employees who report to a single supervisor is almost always unknown, and there is typically no empirical justification for the assumption that it is normal and centered around the scale midpoint (Bernardin & Beatty, 1984). Rather, organizations exert considerable effort to assure that the distribution of performance is not normal. Saal et al. (1980) point out that a variety of activities,

ranging from personnel selection to training, are designed to produce a skewed distribution of performance. Any rational personnel manager would be pleased if all of the employees in a company were exceptional workers. Second, these assumptions imply that there is no variation, from work group to work group, in terms of their actual performance (Murphy & Balzer, 1989). Thus, if I give ratings whose mean is 5.0 (7-point scale), and you give ratings whose mean is 4.1, I am likely to be labeled as the more lenient rater. However, it is entirely possible that my subordinates *are* better performers than yours. There is a substantial literature on leadership that assumes that work groups will differ in their performance, depending in part on the effectiveness of the leader (Landy, 1985). It seems illogical to assume that all groups perform at the same level, regardless of their resources, their leadership, their task, and so on.

The distribution of performance ratings is probably more informative when analyzing the performance rating system of an organization or of some large unit in that organization than when it is used to analyze the ratings given by an individual supervisor. If you find (as is often reported in organizations) that 90% of employees receive ratings of "Well Above Average" or higher, it is likely that your organization will have problems if it tries to use performance ratings to make promotions or to assign merit raises. The tendency of raters to give all of their subordinates high ratings, or even similar ratings, could greatly diminish the potential value of performance ratings for making decisions about promotions, raises, or even training needs (Barrett, 1966; Bretz, Milkovich, & Read, 1992). On the other hand, when we look at the ratings given by individual raters, it sometimes is plausible that the distributions of performance ratings *do* reflect the true performance of employees. Going back to Figure 11.1, it is possible (but perhaps not likely) that Al's subordinates really *are* very good performers. Despite the fact that distributional measures have been used in many studies to indicate the presence and severity of rater errors (DeNisi & Murphy, 2017), our inability to sort out the roles of raters, ratees, and the rating environment in explaining why performance ratings show particular distributions severely limits the value of measures of leniency, severity, central tendency, or range restriction as criteria for evaluating performance ratings.

On the other hand, it is likely that some raters are more lenient in their performance evaluations than others. It has long been believed that some raters are consistently more likely to give high ratings than others (Guilford, 1954), and there is evidence to support this belief (Kane, Bernardin, Villanova, & Peyrefitte, 1995). On the whole, leniency appears to be related to the rater's personality, particularly his or her level of agreeableness and conscientiousness (Bernardin, Cooke, & Villanova, 2000; Roch, Ayman, Newhouse, & Harris, 2005). However, there is also considerable evidence that leniency is influenced by contextual variables such as the purpose of rating. Ratings are consistently highest when they are used for administrative purposes (Bernardin & Orban, 1990; Dobbins, Cardy, & Truxillo, 1986; Harris, Smith, & Champagne, 1995; Murphy & Cleveland, 1995; Taylor & Wherry, 1951). Jawahar and Williams' (1997) meta-analysis suggests that ratings for administrative purposes are one-third of a standard deviation higher than ratings used for research or developmental purposes.

Correlational Errors

Raters tend to give similar evaluations to separate aspects of a person's performance, even when performance dimensions are clearly distinct (Bingham, 1939; Newcomb, 1931; Thorndike,

1920). The result is an inflation of the intercorrelations among dimensions, which is referred to as halo error. Cooper (1981b) suggests that halo is likely to be present in virtually every type of rating instrument.

There is an extensive body of research examining halo errors in rating, and a number of different measures, definitions, and models of halo error have been proposed (Balzer & Sulsky, 1992; Cooper, 1981a, 1981b; Lance et al., 1994; Murphy & Anhalt, 1992; Murphy, Jako, & Anhalt, 1993; Nathan & Tippins, 1989; Solomonson & Lance, 1997). While there is disagreement regarding some minor points, there is clear agreement regarding several propositions:

1. The correlation between ratings of separate performance dimensions reflects both actual consistencies in performance (true halo, or the actual level of correlation between two conceptually distinct performance dimensions) and errors in processing information about ratees or in translating that information into performance ratings (illusory halo).[6]
2. Illusory halo is driven in large part by raters' tendency to rely on general impressions and global evaluations (e.g., Sam is a generally good worker) when rating specific aspects of performance (see, for example, Balzer & Sulsky, 1992; Jennings, Palmer, & Thomas, 2004; Lance et al., 1994; Murphy & Anhalt, 1992).
3. Halo is likely to limit the accuracy of ratings as tools for identifying individual strengths and weaknesses, but it may *enhance* the accuracy of ratings as tools for distinguishing between ratees in terms of their overall performance levels.

This last point might seem counterintuitive, because it implies that rater errors can make performance ratings more accurate. As we will describe in more detail in the section that follows, "accuracy" can mean many different things, and it is in fact true that if ratings on different aspects of performance are highly correlated, this will tend to increase the reliability of overall performance evaluations, in effect maximizing the differences *between* rates by minimizing the differences in ratings *within* rates.

Like distributional measures, measures of halo error are very difficult to interpret because it is difficult, if not impossible, to separate true halo from illusory halo. For example, suppose ratings of Oral Communication are correlated .60 with ratings of Planning and Organization. Does this indicate that halo error is present? This conclusion can only be drawn if there is a good reason to believe that accurate ratings of this set of ratees would result in a different (almost certainly lower) correlation between Oral Communication and Planning and Organization. Even if the expected correlation between two rating dimensions is known in general (for example, in the population as a whole several of the Big Five personality dimensions are believed to be essentially uncorrelated), that does not mean that the performance of a small group of ratees on these dimensions will show the same pattern of true independence.

Spotlight 11.1 Is Halo Error Really an Error?

Bingham (1939) introduced the idea that raters might make an error in evaluating different aspects of the performance of the people they supervised by relying too heavily on general impressions or overall evaluations—that is, they might make halo errors. For the next 50 years,

measures of halo represented one of the most common criteria for evaluating performance ratings. Different researchers suggested different indices, but in general, the higher the correlation among ratings of supposedly separate performance dimensions, the more likely ratings would be regarded as exhibiting halo.

Measures of halo error have always been problematic. As Bingham (1939) recognized, it is very difficult to determine whether ratings of separate aspects of performance are correlated because of halo errors, or whether they are correlated because the dimensions themselves are correlated (Murphy & Anhalt, 1992). Suppose, for example, that when you analyze performance ratings in your organization, you find that ratings of Planning are correlated .50 with ratings of Oral Communication. This might be evidence of halo error, but it also might be evidence that people who are good at planning are also good at oral communication. It is even possible that the correlation between ratings of these two performance dimensions are too low, and that the correlation between the behaviors that constitute Planning and Oral Communication is higher than .50. Murphy, Jako, and Anhalt (1993) argue that there were so many ambiguities associated with the most common measures of halo that they should not be used as criteria for evaluating ratings.

It is worth asking the more basic question of whether we should think of halo as an error at all. First, there is evidence that the validity of common predictors of job performance is *higher* when ratings of separate aspects of performance are highly intercorrelated than when they are uncorrelated (Nathan & Tippins, 1989). Second, the argument that evaluations of separate aspects of performance should be independent of your evaluation of each ratee's overall performance is in many respects illogical (Murphy, 1982b). Would it be possible, for example, for someone to be a poor performer on each performance dimension that is rated, but still be regarded as a generally effective employee? Overall performance is almost certainly based on an employee's success in carrying out the major aspects of his or her job, and in a performance appraisal system that has any level of job-relatedness, the performance dimensions that are rated should cover most of the essential functions of the job. From this perspective, finding that ratings of each of the major aspects of job performance were independent of evaluations of the employee's overall performance would be a signal that there was something wrong with the performance appraisal system.

You might argue that raters commit halo errors if they rely too heavily on their overall impressions when rating separate performance dimensions, but nobody has been able to articulate sensible criteria for determining when raters are paying too much attention, too little attention, or just the right amount of attention to employees' overall effectiveness when rating their performance on specific aspects or dimensions of performance. There is no doubt that raters sometimes make halo errors, but attempts to control or eliminate halo in performance ratings have rarely made things better, and have often had a negative impact on the quality of rating data (Murphy, 1982b; Murphy et al., 1993).

Evaluating Rater Error Measures

As we have noted in several chapters, it is common for the great majority of employees to receive ratings near the top end of the rating scale. It is also common for ratings on conceptually

independent aspects of job performance to be highly correlated. Unfortunately, rater error measures that are based on the distributions and the intercorrelations among the ratings given by an individual rater have proved essentially useless for evaluating performance ratings (DeNisi & Murphy, 2017; Murphy & Balzer, 1989). First, we cannot say with any confidence that a particular supervisor's ratings are too high or too highly intercorrelated unless we know a good deal about the true level of performance, and if we knew this, we would not need supervisory performance ratings. Second, the label "rater error" is misleading. It is far from clear that supervisors who give their subordinates high ratings are making a mistake. Chapters 9 and 12 examine in detail the reasons for rating inflation and related phenomena, and they make the case that raters who give high performance ratings are sometimes making a smart decision about the best way to use performance appraisal systems to motivate their subordinates.

Rater error measures represent one of the first concerted efforts to evaluate performance ratings. Research on so-called rater errors may not have told us much about the quality of rating data, but it has substantially advanced our understanding of topics ranging from cognition to motivation. Chapters 13–14 will discuss some of the insights this research has produced.

Rating Accuracy

Rater error measures and measures of the reliability of ratings can be thought of as indirect assessments of the accuracy of performance ratings. That is, performance ratings that are not reliable cannot be very accurate measures of performance. If ratings reflect little more than random measurement error, they cannot tell you much about the performance of the people being evaluated. Similarly, ratings that are unduly lenient, or that fail to provide distinct information about distinct aspects of performance, are probably not very accurate. Both reliability and rater error measures are essentially negative indicators, in the sense that the absence of measurement error or of distributional or correlational errors makes it possible that ratings are accurate. However, there is no guarantee that error measures that do not suffer from excessive measurement error, leniency, or halo *will* provide accurate assessments of performance. It is possible that raters whose ratings are normally distributed and not highly intercorrelated are nevertheless inaccurate in their assessments of ratee performance.

For many decades, the lack of direct measures of the accuracy of performance ratings was a serious impediment to determining whether or not the conditions under which performance ratings were sufficiently accurate to provide credible assessments of job performance. Starting in the late 1970s, laboratory studies of performance rating provided opportunities for directly measuring rating accuracy. Two distinct types of accuracy measures were developed: behavior-based measures and judgmental measures. Behavior-based measures are considerably simpler, and are based on the rater's accuracy in recognizing specific behavioral incidents (Lord, 1985). For example, Murphy, Philbin, and Adams (1989) studied the effects of purpose of observation on behavior-recognition accuracy. They asked raters to indicate whether each of 36 behavioral incidents had or had not occurred in the videotapes of performance they had observed. Since the true status of each behavior can be determined (here, 18 behaviors actually occurred and 18 did not), it was possible to measure observational accuracy in terms of true positives and true negatives (hits), false positives (false alarms), and false negatives (misses). When responses are

coded in this way (e.g., true positives, true negatives) signal detection theory can be employed to derive measures of response bias and sensitivity (Lord, 1985).

Even more useful were the measures developed by Borman (1977, 1978, 1979). Borman's key insight was that in laboratory studies, where it is common for subjects to make judgments about videotapes of people performing various jobs, it would be possible to develop a credible standard for evaluating whether or not a particular rater's judgments were accurate. In particular, if multiple well-trained raters were given opportunities to view performance under optimal rating conditions (e.g., no distractions of competing task demands), the average of their ratings could be considered a type of "true-score" measure of performance. The true-score measures represent assessments that remove the biases that might characterize individual judgments (by pooling over multiple raters) as well as the errors in observation, encoding, and memory that are typical of performance judgments in the field, and a rater whose judgments are close to those true scores can reasonably be labeled as an accurate rater. Measures of accuracy in judgment have been widely used, especially in cognitively oriented research on performance judgments (Becker & Cardy, 1986; Borman, 1977, 1978, 1979; Cardy & Dobbins, 1986; McIntyre, Smith, & Hassett, 1984; Murphy & Balzer, 1986; Murphy, Balzer, Kellam, & Armstrong, 1984; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982; Pulakos, 1986; see Sulsky & Balzer, 1988 for a review of accuracy measures).

Operational Definitions

Suppose a panel of experts evaluates the performance of 10 individuals, rating each individual on five different performance dimensions. The average rating from this panel could be used as true-score measures, and a comparison between an individual rater and this true-score average could be used to evaluate that rater's accuracy. As Cronbach (1955) noted, there are several ways we could define and assess the agreement between an individual rater who evaluates the same 10 individuals and these true scores. He identified four separate components of accuracy: (1) Elevation, (2) Differential Elevation, (3) Stereotype Accuracy, and (4) Differential Accuracy.

Elevation refers to the accuracy of the average rating, over all ratees and dimensions. So, if the average true score for these 10 individuals is 3.5, and the average rating a particular rater assigns is also 3.5, you could think about this rater as showing perfect accuracy, at least with regard to the overall rating level. Measures of Elevation are conceptually related to leniency measures, with the advantage of providing a direct standard for judging whether a particular rater is giving ratings that are too high (lenient), too low (severe), or in the right ballpark.

Differential Elevation refers to accuracy in discriminating among ratees. If the pooled judgment of experts indicates, for example, that Joe is a better performer than Adam, and that Adam is a better performer than Scott, a rater who also gives higher ratings to Joe and lower ratings to Scott will be accurate in making these between-ratee distinctions. As we noted in Chapter 8, many of the most important uses of ratings (e.g., for making decisions about promotion of salary) depend on making accurate distinctions between people, which suggests that Differential Elevation is a critically important aspect of rating accuracy.

Stereotype Accuracy refers to accuracy in discriminating among performance dimensions. For example, if dimensions being rated include "Planning" and "Oral Communication," Stereotype Accuracy involves accuracy in determining whether a group of workers is in fact better at planning or at oral communication. If the purpose of rating includes diagnosing training and development needs in a work group or an organization, Stereotype Accuracy might be the most important aspect of performance.

Finally, Differential Accuracy refers to accuracy in detecting ratee differences in patterns of performance (i.e., accuracy in diagnosing individual strengths and weakness). For example, in evaluating individual training needs, it might be important to figure out each employee's distinct strengths and weaknesses, and Differential Accuracy might be very important for this purpose.

Cronbach's (1955) found components are not the only measures of the accuracy of performance judgments. Measures of distance accuracy and correlational accuracy have also been proposed (Becker & Cardy, 1986; Borman, 1977; Wiggins, 1973). Research suggests that the different accuracy measures are not highly correlated (Sulsky & Balzer, 1988). For example, Murphy and Balzer (1989) noted that the average correlation among Cronbach's (1955) four measures is essentially zero. One implication is that the conclusions you draw about a rater's accuracy results may depend more on the choice of accuracy measures than on the rater's ability to evaluate his or her subordinates (Becker & Cardy, 1986).

Rater Errors and Rating Accuracy

Research on the relationship between rater errors and rating accuracy confirmed many of the concerns of researchers and practitioners who had questioned the value of using measures of leniency, halo, or other rater errors to evaluate performance ratings. The general trend in this literature is clear: rater error measures are largely unrelated to direct measures of the accuracy of ratings (Becker & Cardy, 1986; Bernardin & Pence, 1980; Borman, 1977; Murphy & Balzer, 1989).

One additional indication of the dubious relationship between rater errors and rating accuracy comes from the rater training literature. A favorite method of training has been to inform raters of the existence and nature of rater errors, and exhort them to avoid those errors. This method does indeed reduce rater errors, but it also reduces the accuracy of ratings (Bernardin & Beatty, 1984; Bernardin & Pence, 1980; Borman, 1979; Landy & Farr, 1983). It appears that rater error training leads raters to substitute an invalid rating bias (avoid rater errors) for whatever strategy they were using before rating. Avoiding errors simply doesn't address the question of accuracy.

Is Accuracy a Useful Criterion?

One of the challenges in evaluating the quality of rating data is that direct measures of rating accuracy are notoriously difficult to obtain in the field (Murphy & Cleveland, 1995). Proxies, such as rater error measures, are not good indicators of accuracy (Murphy, Jako, & Anhalt, 1993). Furthermore, raters appear to have a hard time evaluating the accuracy of the ratings they provide (Roch, McNall, & Caputo, 2011), although they do have some ability to determine their accuracy in observing and recording behavioral information.

There is evidence that we can increase rating accuracy by encouraging raters to take careful behavioral notes and to attend carefully to ratee performance (Mero, Motowidlo, & Anna, 2003; Sanchez & De La Torre, 1996). There is also evidence that increasing accountability (e.g., by asking raters to justify their evaluations) may increase rating accuracy (Mero & Motowidlo, 1995). However, as with many other suggestions for improving the quality of rating data, it is always worth asking whether the probable payoff offsets the probable costs.

Mero et al. (2003) based their conclusions on simulations and laboratory studies, where raters have little to do other than observing and evaluating ratee behavior. In work settings, supervisors and managers have a number of tasks they must perform, and time they spend on taking behavioral notes and careful observation is arguably time they could have spent doing their other tasks. That is, if raters increase the time and energy they devote to getting the most accurate ratings, they will tend to devote less time and energy to their other duties, and this will have a payoff to organizations if and only if getting more accurate ratings is more valuable and useful than whatever else supervisors and managers were doing.

On the whole, research on rating accuracy has been useful, but it has also been limited in terms of the types of questions it can answer. Accuracy measures are only feasible in artificial laboratory environments, where expert ratings obtained under ideal rating conditions can be pooled to create credible "true score" measures. This means that rating accuracy cannot be measured or studied in real-world settings, where there may be barriers to accuracy in rating due to the pressure of competing job responsibilities (unlike laboratory experiments, raters in organizations have a number of jobs to do in addition to observing and evaluating ratee performance) or to a wide range of motivational factors that will be discussed in Chapter 12. While accurate ratings might be desirable in many situations, there are simply no good methods for evaluating the accuracy of ratings in most organizational settings. Instead of directly evaluating accuracy, we are often forced to rely on indirect criteria. Some of these, such as measures of agreement and reliability, are broadly useful, while others, such as measures of halo or leniency, are of dubious value. An alternative to relying on measures of agreement and rater error measures to evaluate the quality of rating data is to apply a strategy that is widely used in psychological testing—that is, using multiple types of data to evaluate the construct validity of performance ratings (Astin, 1964; James, 1973; Smith, 1976).

Construct Validity of Performance Ratings

Many tests and assessments are designed to measure *constructs*, such as Intelligence or Agreeableness. A construct is a label we use to describe a set of related behaviors or phenomena. Constructs do not exist in a literal sense (you cannot put a pound of Agreeableness on the table in front of you), but they are an extremely useful tool for describing and understanding human behavior. Most assessments of the validity of tests, rating scales, and the like can be thought of as assessments of the construct validity of those instruments (Cronbach, 1990; Murphy & Davidshofer, 2005).

There is no single method that is used to evaluate the construct validity of a test or a measure. Rather, construct validation involves the collection of evidence from a range of sources about whether a test measures what it is intended to measure. In this sense, the assessment of the

construct validity of performance ratings involves collecting evidence that can help you determine whether or not, and how well, performance ratings measure actual job performance.

Several classes of evidence might be used in evaluating the construct validity of job performance ratings. First, we might look at the content of the scales used to rate performance. Rating scales that are based on a careful analysis of the job and that provide clear, unambiguous definitions of the performance dimensions to be rated are more likely to provide a valid measure of job performance than ambiguous scales that are not clearly job related. In Chapter 6, we noted that one advantage of using behaviorally anchored rating scales was that these scales provided clear and concrete definitions of the performance dimensions to be rated. This clarity could almost certainly contribute to the construct validity of performance ratings.

Second, we might look at evidence of convergent validity—that is, the extent to which performance ratings and other measures of job performance provide consistent information. For example, there is evidence that performance ratings are positively correlated with a number of objective measures of job performance. Several reviews have suggested that performance ratings and objective performance measures are related (with corrected correlations in the .30s and .40s), but they are not substitutable (Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995; Conway, Lombardo, & Sanders, 2001; Heneman, 1986; Mabe & West, 1982).

Third, we might look at criterion-related validity evidence. Performance ratings are one of the most common criteria for validating personnel selection tests, and there is a massive body of research showing that tests designed to measure job-related abilities and skills are consistently correlated with ratings of job performance (Schmidt & Hunter, 1998, review 85 years of research on the validity of selection tests, and document consistent correlations between many predictors and job performance ratings; see also Woehr & Roch, 2016). Usually, we think of these correlations as evidence for the validity of selection tests, but we can just as well use them as evidence for the construct validity of performance ratings. That is, there is a substantial body of evidence showing that measures (e.g., ability tests) that *should be* related to job performance *are*, in fact, related to ratings of job performance. This suggests that performance ratings are capturing at least something about the performance of those individuals being evaluated.

Another way of gathering evidence about the construct validity of performance ratings is to determine whether ratings have consistent meanings across contexts or cultures. Performance appraisals are used in numerous countries and cultures; multinational corporations might use similar appraisal systems in many nations. The question of whether performance appraisal provides measures that can reasonably be compared across borders is therefore an important one. Ployhart, Wiechmann, Schmitt, Sacco, and Rogg (2003) examined ratings of technical proficiency, customer service, and teamwork given to fast-food workers in Canada, South Korea, and Spain and concluded that ratings show evidence of rating invariance. In particular, raters appear to interpret the three dimensions in similar ways and to apply comparable performance standards when evaluating their subordinates. However, there was also evidence of some subtle differences in perceptions that could make direct comparisons across countries complex. In particular, raters in Canada perceived smaller relationships between Customer Service and Teamwork than did raters in South Korea and Spain. On the whole, however, Ployhart et al. (2003) concluded that ratings from these countries reflected similar ideas about the dimensions

and about the performance levels expected, and could therefore be used to make cross-cultural comparisons.

Similarly, there is evidence of measurement equivalence when performance ratings of more experienced and less experienced raters are compared (Greguras, 2005). Even though experience as a supervisor is likely to influence the strategies different supervisors apply to maximize the success of their subordinates, it appears that supervisors using a well-developed performance appraisal system are likely to agree regarding the meaning of performance dimensions and performance levels.

Finally, we might use evidence regarding bias in performance ratings to evaluate the construct validity of these ratings. The rationale here is that if performance ratings can be shown to be strongly influenced by factors other than job performance, that would tend to argue *against* the proposition that performance ratings provide valid measures of job performance (Colella, DeNisi, & Varma, 1998). There is a substantial literature dealing with the question of whether or not performance ratings are biased by factors that are presumably unrelated to actual job performance, such as the demographic characteristics of ratees or the characteristics of work groups.

Are Performance Ratings Biased?

There are a number of reasons to believe that performance ratings might be biased. Several theories and models in social and cognitive psychology, such as role congruity theory (Nieva & Gutek, 1980), the lack of fit model (Heilman, 1983), relational demography theory (Tsui & Gutek, 1999), and stereotype fit models (Dipboye, 1985), predict that attributes such as race, gender, or age *should* influence supervisors' perceptions of the competence and performance of their subordinates, particularly subordinates who are members of minority groups. Roberson, Galvin, and Charles (2007) note that despite some progress in workplace equity, there are still large differences in the likelihood that women or members of racial and ethnic minority groups will advance to the top of their organizations, and they suggest that bias in performance appraisal might be a factor in these differences.

There are some studies showing what appears to be bias in performance ratings. For example, laboratory studies suggest that women may receive lower evaluations when they occupy jobs that are stereotypically held by men (Lyness & Heilman, 2006). Other studies suggest that older raters tend to rate older employees a bit more favorably, while younger raters tend to favor other younger employees (e.g., Gibson, Zerbe, & Franken, 1993). It is also likely that some attributes are viewed unfavorably by most of the population (e.g., obesity, low levels of physical attractiveness), and these broadly negative characteristics could lead to lower ratings (Bento, White, & Zacur, 2012; Nieminen et al., 2013).

Studies of demographic biases in performance rating sometimes appear to yield mixed results. Table 11.3 summarizes the results of 19 studies of the relationship between age, gender, race, and disability and performance ratings. This table glosses over numerous important features of many of these studies (e.g., different conditions under which ratings were obtained), including variability in what is being rated (e.g., overall performance, trainability), and it is far from comprehensive (no studies prior

to 1993 are included). Nevertheless, this table does suggest that the literature on potential biasing factors in performance appraisal might yield mixed results.

A much clearer picture of the influence of demographic variables on performance rating emerges from the numerous meta-analyses that have been conducted in this field. Table 11.4 summarizes the key findings of these analyses. The values shown in Table 11.4 are based on the combined results of many individual studies, and they are remarkably consistent. On the whole, they suggest that age, gender, race, and disability tend to have small effects on performance ratings (Landy, 2010). There are no doubt specific conditions where these variables might have larger effects (e.g., Heilman & Chen [2005] report larger differences when men and women perform helping behaviors), but the hypothesis that performance ratings are substantially biased against women, members of minority groups, older workers, or disabled workers does not seem credible.

Table 11.3 Results of 19 Studies of Demographic Differences in Performance Ratings

Little or No Effect	**Meaningful Difference in Ratings**
Age	
Cox & Beier (2014)	Finkelstein, Burke, & Raju (1995)
Lefkowitz & Battista (1995)	Rupp, Vodanovich, & Credé (2006)
Ng & Feldman (2008)	
Siegel (1993)	
Treadway, Ferris, Hochwater, Perrewé, Witt, & Goodman (2005)	
Gender	
Bowen, Swim, & Jacobs (2000)	Lefkowitz & Battista (1995)
Lyness & Heilman (2006)	Heilman, Block, & Stathatos (1997)
Robbins & DeNisi (1993)	Heilman & Chen (2005)
Race/ethnicity	
Baltes, Bauer, & Frensch (2007)	Ho, Thomsen, & Sidanius (2009)
McKay & McDaniel (2006)	
Roth, Huffcutt, & Bobko (2003)	
Disability	
Colella, DeNisi, & Varma (1998)	
Miller & Werner (2005)	
Ren, Paetzold, & Colella (2008)	

Several other authors (e.g., Arvey & Murphy, 1998; Bass & Turner, 1973; Baxter, 2012; Bowen et al., 2000; DeNisi & Murphy, 2017; Kraiger & Ford, 1985; Landy, Shankster, & Kohler, 1994; Pulakos, White, Oppler, & Borman, 1989; Waldman & Avolio, 1991) have reached similar

conclusions regarding the lack of bias in performance ratings, as have studies of more specific biases (e.g., pregnancy bias; Gueutal, Luciano, & Michaels, 1995, suggest that pregnancy has a small effect on ratings, and that pregnant workers receive higher ratings) but this conclusion is not universally accepted. Stauffer and Buckley (2005) note that while studies of racial and ethnic differences in performance ratings do show relatively small mean differences, there are some important caveats. In particular, they present data showing that there are nontrivial differences in the ratings received by white and black ratees when the rater is white (when the rater is white, white ratees receive higher ratings). However, even in these carefully selected rater–ratee pairs, the proportion of variance in performance ratings explained by race is fairly small (approximately 2.5% of the variance in ratings performed by white raters is explained by rate race).

Table 11.4 Meta-Analytic Estimates of Biasing Factors on Performance Appraisal

Potential Biasing Factor	Percentage of Variance in Overall Performance Ratings Explained
Age	Less than 1%[a]
Gender	Less than 1%[b]
Race & Ethnicity	1.7%[c,d]
Disability	1.5%[e]

[a] Ng & Feldman (2008); [b] Bowen, Swim, & Jacobs (2000); [c] McKay & McDaniel (2006); [d] values presented in Roth, Huffcutt, & Bobko (2003) are corrected for attenuation, but the value shown here is uncorrected. Both McKay & McDaniel (2006) and Roth, Huffcutt, & Bobko (2003) reported average uncorrected *d* values of .27, which translated into r^2 = .017; [e] Ren, Paetzold, & Colella (2008).

Studies using laboratory methods (e.g., Hamner, Kim, Baird, & Bigoness, 1974; Rosen & Jerdee, 1976; Schmitt & Lappin, 1980) sometimes suggest that there are larger demographic differences in performance ratings. However, there are reasons to believe that the findings of these laboratory studies (particularly those involving vignettes rather than observations of actual performance) overestimate the effects of the demographic characteristics of both raters and ratees (Landy, 2010). Wendelken and Inn (1981) noted that demographic differences are made especially salient in laboratory studies where other ratee characteristics (including performance levels) are tightly controlled and where raters are untrained and have no prior knowledge of and no relationship with ratees. Murphy, Herr, Lockhart, and Maguire's (1986) meta-analysis confirmed that vignette studies do indeed tend to produce larger effects.

Most assessments of bias in performance appraisals focus on the numerical ratings that are provided. Wilson (2010) went further, examining potential differences in the narrative comments that are included in most performance appraisals. He hypothesized that members of minority groups might receive more negative comments, but he found small differences in the opposite direction. On the whole, supervisors made positive comments, even when ratees received lower ratings. Asian employees were slightly more likely to receive positive comments than white employees, while black employees were slightly more likely to receive negative comments than their white counterparts. The operative term here is "slightly"; on the whole, there are few differences in the narrative comments received by ratees from different racial/ethnic groups.

Finally, it is often argued that performance appraisal systems that rely on general rather than specific performance criteria (e.g., graphic scales versus behavioral scales) are more prone to

bias and unfair discrimination; expert witnesses in employment discrimination litigation often testify to this effect. In fact, there is little evidence to support this belief (Bernardin, Hennessey, & Peyrefitte, 1995; Hennessey & Bernardin, 2003). As noted above, demographic differences in ratings are generally small, and what differences there are do not seem to be affected by the rating scales used.

Conclusions About the Reliability, Validity, and Accuracy of Performance Ratings

Performance appraisal researchers have spent decades developing methods and indices that can be used to evaluate the quality of rating data. Some methods have proved more useful than others and some problems have turned out to be easier to solve than others. Our review of this research leads to four broad conclusions, shown in Table 11.5.

Table 11.5 Four Conclusions About the Reliability, Validity, and Accuracy of Performance Ratings

Ratings of performance are determined in part by the behavior and effectiveness of the employees being rated and in part by factors that seem to have little to do with performance.
Performance ratings are influenced by both random and systematic measurement errors, and the reliability of performance ratings is neither as high as the reliability of well-constructed tests and assessments nor as low as most researchers and practitioners assume.
The measures most frequently used to evaluate the quality of rating data (i.e., rater error measures) are probably the least useful for this purpose.
It is easier to evaluate the quality of rating data in the aggregate (e.g., at the organizational level) than it is to evaluate the quality of rating data provided by a single rater.

First, it seems clear that performance ratings tell you *something* about the performance of the individuals being rated, but that performance ratings are influenced by things that have little to do with job performance (Landy & Farr, 1980; Milkovich & Wigdor, 1991). By the 1970s, it was pretty well established that performance ratings showed evidence of reliability and validity, but it was also well understood that many factors other than performance were important in determining the ratings an individual rater received. Research in the 1980s examined the role of cognitive processes in determining the strengths and weaknesses of performance ratings; this research showed how the way we attend to, mentally represent, and recall behavior could influence ratings. By the 1990s, researchers were starting to turn their attention to the role of contextual factors, ranging from broad national and organizational contexts to the immediate environment in which performance is observed and evaluated influenced ratings, and this contextual orientation is still a dominant theme in research on performance ratings. On the whole, all of these streams of research have shown that it is unrealistic to expect performance ratings to be a perfect representation of the performance of the individuals being rated, but that these ratings can and do provide useful information about performance.

Second, measurement error is a serious problem in performance rating, and it is important to devise practical strategies for reducing the influence of measurement error. On the other hand, the widely circulated claim that supervisory ratings represent an almost equal mix of valid information about ratee performance and random measurement error (Viswesvaran, Ones, &

Schmidt, 1996) is clearly wrong. Researchers who have applied classic test theory to the analysis of performance ratings have reached two conclusions—that ratings are highly unreliable and that it is easy to correct for the effects of this unreliability, which do not hold up when more comprehensive models of measurement (e.g., generalizability theory) are applied to the same data. It now seems clear that there are both systematic and random sources of measurement error in performance ratings, and that the influences of both of these depends very much on what one is trying to capture with ratings. Thus, the reliability of information about between-person differences might be quite different from the reliability of information about within-person differences. Performance ratings can be highly reliable for some purposes and less reliable for others.

Third, the most widely used criteria for evaluating ratings—rater error measures—are sometimes the least useful. As we noted in our discussion of rater error measures, it is often hard to tell whether ratings are so high that you can conclude that a particular rater is lenient or so intercorrelated that you can conclude that halo error is present. When you use measures of this sort to evaluate the ratings provided by an individual rater, it is hard to tell whether or to what extent the distributions or intercorrelations those ratings exhibit is a reflection of the performance being evaluated or the actual behavior of the employees you are evaluating.

Fourth, the task of evaluating ratings is more challenging when trying to draw conclusions about the ratings obtained from a particular rater than when trying to draw conclusions at some higher level of analysis. For example, if the average rating I give to subordinates is 4.25 on a 5-point scale, this *might* indicate that I am lenient, but it could also reflect the truly above-average performance of my subordinates. On the other hand, if the average rating in an entire organization is 4.25 on a 5-point scale, it is hard to accept the possibility that *everyone* is above average, and it is more likely that the ratings in this organization are unrealistically high. This is exactly what many assessments of organizational-level leniency have found; Chapters 9 and 12 explore the reasons why performance ratings are so often unrealistically high. Traditional measures of leniency and halo were developed with the individual rater in mind, but they might prove more valuable when evaluating ratings in the aggregate than when trying to draw conclusions about individual raters.

Are Ratings Useful?

Performance appraisal researchers have spent a long time asking whether performance ratings are reliable, valid, and accurate, and devising methods of measuring these attributes. A case can be made that we have spent a lot of time asking the wrong question, and that it is more important to ask whether ratings are *useful*. In the best of all possible worlds, organizations would use performance ratings for a range of important purposes, starting with using ratings to help make decisions about important rewards and sanctions (e.g., raises, promotions, identifying candidates for layoffs), to help direct employee training and development, to evaluate and validate HR policies (e.g., to determine whether selection tests are likely to lead to better hiring decisions), and to draw conclusions about the health of the organization. In some organizations, it is not clear that *any* of these purposes are actually achieved, and we would argue that if organizations are not going to use performance ratings for purposes of this sort they should not collect them.

Validity and accuracy might be a precondition for the optimal use of ratings. That is, if an organization truly wants to distribute rewards on the basis of job performance, or to tailor training to the strengths and weaknesses of individual employees, reliable, valid, and credible information about performance is a starting point. However, accurate performance ratings are no guarantee that they will be used effectively, and as we will see in Chapters 9 and 12, accurate performance ratings can cause more problems than they solve. For example, we noted in Chapter 5 that people tend to evaluate their own performance more positively than other people evaluate it. One implication is that truly accurate performance feedback is likely to *feel* to the recipient like it is unfair and unduly negative; we examined the implications of this in Chapters 9 and 10. We believe it is absolutely critical to understand how organizations actually use performance ratings (if they use them at all in any meaningful way) in order to determine whether they are useful. As we noted in Chapter 8, organizations often shoot themselves in the foot by trying to use the same performance appraisal system for multiple conflicting purposes. The most comprehensive evaluation of the usefulness of ratings must go beyond looking at the ratings themselves, to consider what the organization is trying to accomplish with performance appraisal and to identify (and if possible remove) the barriers to accomplishing these goals.

Summary

A number of approaches have been identified for evaluating the quality and potential usefulness of performance ratings. First, we might consider whether different raters who are evaluating the same employees reach similar conclusions. If they do not, we may need to make hard decisions about whose ratings to believe and whose to give the most weight to. Second, we might examine the distributions and the intercorrelations among ratings. If performance ratings are seriously skewed (e.g., if most employees are rated well above average) or if ratings on conceptually distinct areas of performance are redundant, that might limit the potential value or uses of rating data. Measures of so-called rater errors are probably more useful for evaluating ratings across organizations or across many raters, and are harder to interpret when used to evaluate an individual rater. Similarly, assessments of the validity of ratings appear more useful for aggregated ratings than for drawing conclusions about the ratings given by an individual supervisor or manager.

In laboratory settings, it is possible to develop measures of the accuracy of performance ratings, and while these have some value for performance appraisal research, they have only limited applicability in the field. These measures have proved useful for exploring the factors that contribute to or detract from accuracy, but they are not practical for use in organizations.

The ultimate criterion for evaluating rating data is the usefulness of ratings, and usefulness depends as much on what the organization does or attempts to do with performance ratings as on the ratings themselves. To be sure, criteria such as reliability, validity, and accuracy have implications for evaluating usefulness—ratings that have little to do with the actual performance of the individuals being evaluated are unlikely to be useful. However, the extent to which performance ratings are useful will necessarily depend on answering the question *"useful for what?"* and the answer to this question is not always clear. Too often performance appraisal seems like an exercise in futility, in the sense that a lot of information is collected (at the cost of large investments of time and other resources) with little apparent purpose. If you cannot

convincingly answer the question of why your organization collects performance ratings, you should seriously consider shutting down your performance appraisal program.

Exercise: Analyze Rating Data

Near the beginning of this chapter, we presented some data in Table 11.1 to illustrate some of the questions that might be asked when evaluating performance ratings. In that table, two raters evaluated the performance of five nurses, rating them on four separate performance dimensions (Quality, Timeliness, Accuracy, Judgment). One way to get a concrete sense of how performance ratings are evaluated is to analyze these data and see what conclusions you might draw.

First, create a spreadsheet with 20 rows and 4 columns, where the first two columns identify the target (who is rated) and the performance dimension being rated and the third and fourth columns include the performance ratings the two raters assigned. If you copy this dataset into a data analysis program (e.g., SPSS, R), you will be able to ask several questions about inter-rater agreement. For example, do raters agree more in their evaluations of some dimensions than others?

If you select one performance dimension at a time and correlate the ratings of the five nurses on each dimension, you should find:

	Inter-Rater Correlation
Quality	.00
Timeliness	.45
Accuracy	.01
Judgment	.31

That is, there is virtually no agreement at all in ratings of Quality or Accuracy, with much higher levels of agreement for ratings of Timeliness and Judgment.

Suppose you wanted to ask whether raters agreed, in general, in terms of their rank-ordering of the nurses' performance. This would require a new data set, with five rows and two columns, with the average rating (across the four performance dimensions) assigned by each rater to each of the five nurses. If you correlate these two ratings, you should find $r = .149$, which indicates a low level of agreement. On the whole, these two raters do not agree in their evaluation of these five nurses.

Finally, suppose you wanted to apply the concepts associated with generalizability theory to get a better sense of which factors explain variations in rating and what conclusions might be drawn about the reliability of these performance ratings. This analysis is most easily done by creating a data set with 40 rows and four columns indicating the target (which nurse is rated), the performance dimension, the rater, and the performance rating that is assigned. This allows you to do a simple Target × Dimension × Rater analysis of variance, which yields:

Source	Sum of Eta Squared[7]	
Target	22.15	.297
Dimension	.40	.005
Rater	2.50	.033
T × D	12.85	.172
T × R	18.75	.252
D × R	8.30	.111
Residual (T × D × R)	9.45	.127
Total	74.39	

Eta Squared indicates the percentage of the total variability in scores due to each of the sources. This analysis suggests that there are two important sources of variability in performance ratings, differences between nurses (Targets) and Target by Rater interactions (i.e., the two raters have different ideas about which nurses perform best and worst). You could use the results of this analysis of variance to calculate generalizability coefficients, and while these can be informative and useful, the descriptive information provided by the Eta Squares is probably even more useful, because it allows you to understand both strengths (there are substantial differences in ratings across the people being rated) and weaknesses (the raters do not agree about which nurses perform best and worst) of the performance ratings shown in Table 11.1.

The data analyzed here would not give you much confidence in these performance ratings. What you hope to find is stronger and more consistent agreement, and also to find that sources of systematic disagreement (e.g., T x R interactions) have only a small effect on ratings. That is not what this particular set of data shows.

1. Their overall ratings are almost completely uncorrelated, with an inter-rater agreement level of $r = .14$.
2. Sam's mean rating is 4.95, Jose's is 5.45.
3. If you calculate the standard deviation in the ratings given to each nurse, then average these, the variability of Jose's ratings of each nurse is 34% larger than the variability of Sam's (1.10 vs .82).
4. If we symbolize the reliability of a measure as r_{xx}, then $1 - r_{xx}$ = the proportion of the variance in ratings due to random measurement error.
5. Some studies, however (e.g., Facteau & Craig, 2001; Maurer, Raju, & Collins, 1998) suggests that ratings from different sources provide reasonably equivalent measures.
6. Bingham (1939) was the first to distinguish between true and illusory halo.
7. Eta squared for targets = Sum of Squares for Targets/Sum of Squares Total = 22.15/74.39.