CHAPTER 9

Reliability, Validity, and Meaningfulness of Multisource Ratings

Kevin R. Murphy
Jeanette N. Cleveland
Carolyn J. Mohler

Multisource rating systems are increasingly popular (Church and Bracken, 1997), especially as a tool for employee development. These systems often include input from one or more supervisors, several peers, and several subordinates, and they may also include self-assessments or evaluations from others within or outside the organization (for instance, customer feedback).

A key assumption of a multisource rating system is that these ratings contain information that is useful and relevant to the individuals being evaluated, and that the ratings are the basis for assessment, training and development, career planning, and other similar activities. The psychometric characteristics of multisource ratings are therefore an important consideration.

Research on the reliability, validity, and other characteristics of ratings can help us understand whether, and under what conditions, multisource ratings are likely to produce information that is potentially useful to the recipient and the organization.

This chapter focuses on three questions. First, do multisource ratings contain consistent and stable information about ratees' performance and behavior in organizations (reliability)? Second, what do these ratings tell us about the individuals being evaluated—for example, do they give information about people's performance and effectiveness on the job (validity)? Third, is this information useful and meaningful to the recipient (meaningfulness)?

Reliability of Multisource Ratings

There is an extensive literature discussing the reliability of performance ratings, much of which focuses on supervisory ratings. For the most part, research on the reliability of supervisory ratings has focused on the stability of overall scores obtained from a single supervisor, who uses a multi-item performance appraisal form (Murphy and Cleveland, 1995). Early meta-analytic studies (for example, Schmidt, Hunter, and Caplan, 1981) used a test-retest approach to estimate the reliability of ratings, but more recent research has focused on interrater reliability estimates. This research has produced findings that are remarkably consistent, and generally discouraging. First, if two raters are asked to evaluate an individual's job performance, their ratings are not likely to be highly correlated. Second, interrater agreement or disagreement does not seem to depend much on organizational level (supervisors, peers, subordinates, etc.). No matter who is asked to evaluate job performance, it is likely that they will disagree.

Agreement Within Sources

Conway and Huffcutt's meta-analysis (1997) reveals that subordinates showed the lowest level of interrater reliability. On average, subordinate ratings of job performance show correlations in the low .30s; average interrater correlations are slightly higher for peers (.37). Supervisors show slightly higher levels of agreement (.50), but again, similarly situated raters tend to provide evaluations that are only moderately consistent. Viswesvaran, Ones, and Schmidt (1996) report similar estimates. In their review, average interrater correlations were .52 for supervisors and .42 for peers.

Peer ratings are sometimes viewed as superior to ratings from other sources, in part because of their supposedly higher reliability. One reason that peer ratings *seem* more reliable than supervisory ratings is that peer ratings are often averaged over several individuals, whereas supervisory ratings are usually obtained from a single

individual (Scullen, 1997). In fact, if you adjust for the effects of this aggregation, peer ratings are probably *less* reliable than ratings obtained from other sources (Viswesvaran, Ones, and Schmidt, 1996). For the most part, the enhanced reliability and validity of peer ratings is a by-product of the fact that they are often presented in aggregated form rather than an indication of peers' enhanced abilities or the shortcomings of supervisors as raters.

Agreement Between Sources

A number of studies have examined agreement in ratings obtained from different sources (say, agreement between supervisory and peer ratings). For example, Harris and Schaubroeck's meta-analysis (1988) reported that an average correlation between peer and supervisor ratings was .53 (this figure is very similar to Viswesvaran and colleagues' estimate of the average correlation between supervisory ratings, namely, .52), the average self-supervisory correlation was .31, and the average self-peer correlation was also .31. They note that there are many reasons why different sources should *not* agree, including the effects of egocentric biases (self-evaluations often differ from evaluations received from others; Thornton, 1980), differences in opportunities to observe (Murphy and Cleveland, 1995), and differences in organizational level.

A more recent meta-analysis (Conway and Huffcutt, 1997) suggested that correlations between ratings obtained from different sources (supervisors, peers, subordinates) are even lower than this. This analysis suggests that the mean correlation between ratings obtained from a supervisor and a single peer is .34, and all other correlations between ratings obtained from different sources are in the .20s or lower. Recent studies of multisource rating systems (Scullen, Mount, and Goff, forthcoming; Greguras and Robie, 1998) confirm this general pattern. Raters typically do not agree, and it does not matter much whether they are at the same level (as with agreement between two supervisors) or different levels in the organization (supervisor-peer correlations).

What Do Interrater Correlations Tell Us?

A number of researchers have suggested that interrater correlations provide an estimate of the reliability of ratings (Schmidt and Hunter, 1996; Viswesvaran, Ones, and Schmidt, 1996). Correlations between ratings from similarly situated raters are rarely much greater than .50, and correlations between ratings from different sources are sometimes even lower. If interrater correlations are interpreted as reliability estimates, this means that 50 percent or more of the variance in performance ratings is probably due to measurement error.

The argument that interrater correlations can be interpreted as reliability coefficients is based on a model that treats raters as passive measurement instruments. For example, Schmidt and Hunter claim that each "rater is analogous to a different form of the rating instrument" (1996, p. 209). In another paper, Viswesvaran, Ones, and Schmidt (1996) argue that the question of interest in evaluating the reliability of performance ratings is whether "the same ratings will be obtained if a different but equally knowledgeable judge rated the same employee" (p. 565). If raters are viewed as alternate forms of a measurement instrument, the correlation between these alternate forms should constitute an estimate of reliability.

There are several reasons to believe that raters in organizations cannot be treated as interchangeable forms of a rating instrument. First, in most organizations, raters observe different behaviors and have differing responsibilities when completing performance ratings (Borman, 1974; Murphy and Cleveland, 1995). Indeed, one explanation for disagreements between raters is that they are not equally knowledgeable (Borman, 1974) but rather observe fundamentally different aspects of a ratee's behavior.

Second, and more important, treating raters as interchangeable measurement instruments implies that measurement is a primary—or at least an important—aspect of performance rating in organizations. Most reviews of performance rating research (Cleveland and Murphy, 1992; Landy and Farr, 1980; Long, 1986; Longenecker, Sims, and Gioia, 1987; Milkovich and Wigdor, 1991; Murphy and Cleveland, 1995) suggest that raters pursue a number of goals when completing performance appraisals (for example, motivating subordinates, maintaining smooth interpersonal relationships), and that accurately evaluating their subordinates is often a relatively minor concern of raters. In contrast to a model in which raters function as alternate forms of a single measurement instrument, this research suggests that performance rating is a complexly

motivated activity, sometimes driven by variables that have little to do with ratees' performance.

Low interrater correlations are usually taken as evidence of random measurement error in ratings, but there is a much simpler explanation for raters' failure to agree. Performance ratings are normally collected in settings where range restriction is ubiquitous, especially when ratings are used to make administrative decisions about ratees (such as salary and promotion; Murphy and Cleveland, 1995). For example, Bretz, Milkovich, and Read (1992, p. 333) conclude that "the norm in U.S. industry is to rate employees at the top end of the scale." Range restriction of the sort that is absolutely routine in performance appraisal can substantially limit interrater agreement. For example, if the .52 correlation cited by Viswesvaran, Ones, and Schmidt (1996) as an estimate of reliability is corrected for the level of range restriction typically found in real-world performance appraisals, reliability estimates are more likely to be in the .70s and .80s than in the .50s.

Even if low interrater correlations are not an indication of low reliability, they do represent a serious problem for multisource rating systems. As we noted earlier, any system that features multiple ratings is likely to produce substantial disagreements between raters. Disagreements might be more pronounced when comparing evaluations from different sources, but even if all ratings are collected from the same source, disagreement is likely to be the norm, not the exception.

One potential solution to the problems of low agreement among raters is to pool data from several raters and provide feedback in the form of aggregated ratings. Aggregation is widely recommended as a potential solution to limited reliability (Scullen, Mount, and Goff, forthcoming; see, however, Greguras and Robie, 1998). This technique might be even more valuable as a means of reducing the inconsistency in ratings and performance feedback that is virtually guaranteed if individual ratings from multiple sources are fed back to employees.

Is Aggregation a Solution or a Problem?

There are two reasons one might want to present multisource ratings in aggregated form (for example, if there are four peer ratings, the average over peers can be presented to the recipient rather than four separate ratings). First, aggregating ratings increases their reliability (Scullen, 1997). In the preceding section, we noted that low interrater agreement was a chronic problem in multisource feedback (MSF) systems, and aggregation can certainly help here. Greguras and Robie (1998) note that MSF programs rarely include sufficient numbers of raters at any level, especially supervisors, to achieve high reliability, so aggregation must be thought of as a partial rather than a full solution to the reliability problem; still, it can help.

A more compelling reason for aggregating is that it may reduce the potentially disruptive influence of interrater disagreements. As noted earlier, multisource rating systems are almost certain to produce inconsistent evaluations, and they might produce more confusion than clarity. The low levels of interrater agreement reviewed earlier mean that many ratees are likely to receive favorable evaluations from some raters and unfavorable ones from others. This may diminish the credibility of ratings and lead to "selective listening," in which unfavorable feedback is dismissed. Aggregated ratings may give a better picture of how an individual's performance is viewed by peers, supervisors, etc., as a group than can be obtained from individual ratings.

One potential downside of the strategy of aggregating ratings is that aggregation may have undesirable effects on the ratings themselves. London, Smither, and Adsit (1997) suggest that accountability may be the Achilles heel of multisource rating systems. They correctly note that raters often have even less accountability under multisource systems (ratings may be anonymous, or averaged over raters) than under traditional systems. Cleveland and Murphy (1992) note that raters are only rarely held accountable by organizations, and the accountability pressures that do exist (for instance, the need to give subordinates disappointing feedback) often serve to inflate ratings. Thus, there may be relatively little accountability to lose in moving from single-rater to multiple-rater scenarios. Nevertheless, the concerns expressed by London, Smither, and Adsit (1997) are important ones. Multisource rating systems may undercut what little accountability exists in rating systems, and this is likely to detract from the quality and usefulness of rating data.

Should Ratings Be Aggregated by Source?

The idea of aggregating ratings by source—for example, averaging peer ratings together—makes intuitive sense, and it is common practice in multisource rating systems (Bozeman, 1997). However, there is little clear evidence that different sources provide truly distinct information, or that aggregating by source really makes sense. For example, Mount and others (1998) suggest that source effects in multisource rating systems are small, and that what are often taken for source differences (disagreements between, for example, peers and supervisors) are probably due to the generally low levels of agreement between raters, regardless of their position in the organization (similar results have been reported by Conway and Huffcutt, 1997; and Scullen, Mount, and Goff, forthcoming). That is, grouping ratings by peers, supervisors, subordinates, etc., may not make much psychometric sense. On the other hand, it makes obvious psychological sense, and it is likely to enhance the acceptability and impact of ratings. It is important to keep in mind that pooling ratings by source may lead to unwarranted conclusions about differences in how your performance is viewed by superiors, peers, subordinates, and so on; nevertheless, aggregating by source is likely to provide more benefits than drawbacks.

Validity of Multisource Ratings

Several methods have been used to estimate the validity of performance ratings. The most obvious approach is to correlate ratings with objective measures of performance and effectiveness. For example, Heneman (1986) reviewed correlations between supervisory ratings and results-oriented measures of performance. The corrected mean correlation between the two was .27, but there was substantial variability in the r values. Higher correlations were found for some rating methods (composite ratings and relative ratings produced higher correlations), but these two classes of measures cannot be treated as interchangeable. A subsequent meta-analysis (Bommer and others, 1995), which focused solely on objective measures of countable behaviors or outcomes, suggested that the correlation between supervisory ratings and objective performance indices was substantially higher (a corrected correlation of .39 was reported). Conway, Lowe, and Langley (1999) report a similar meta-analysis involving peer and subordinate ratings; they report corrected correlations between ratings and objective measures of .34 and .30 for peers and subordinates, respectively. These reviews suggest that performance ratings and objective measures do overlap but are not interchangeable.

Correlations between objective measures and ratings produce useful information, but they are not by themselves a comprehensive index of the validity of ratings. Objective measures of performance and effectiveness rarely capture all of the facets of the performance domain (Murphy and Cleveland, 1995), and it is not clear that objective measures are any better than subjective judgments as an indication of how well or poorly individuals perform. An alternative to relying on correlations with objective measures to evaluate validity is to adopt a construct validation approach. A review of performance ratings conducted by the National Research Council (Milkovich and Wigdor, 1991) concluded that supervisory ratings of performance do indeed show evidence of construct validity. This review did not explicitly consider the validity of peer, subordinate, or self-ratings, but the patterns of evidence that led to the conclusion that supervisory ratings are valid also appears to apply to peer and subordinate ratings.

Woehr, Sheehan, and Bennett's analysis (1999) suggests that supervisors, peers, and other sources agree substantially in terms of the constructs that underlie their ratings, and further that interrater disagreements cannot be dismissed as measurement error. Raters observe different behaviors and apply various standards when evaluating behaviors (see also Murphy and Cleveland, 1995), and their disagreements may in part reflect systematic differences in what they are evaluating. In any case, multisource ratings do appear to show evidence of construct validity.

Validity for What?

Murphy and Cleveland (1995) note that performance ratings cannot be thought of as tests or measurement instruments designed simply to give a numerical estimate of someone's performance. Rather, performance ratings reflect a complex interaction between what the ratee is doing, the goals of the rater (see also Cleveland

and Murphy, 1992), the context in which ratings occur, etc. If we define *validity* in terms of the question "Do these ratings reflect the person's true performance level?" we are likely to come to different conclusions about validity than if we ask another set of questions ("Do these ratings reflect other people's *perceptions* of performance and effectiveness?"). That is, multisource ratings may or may not yield information about "true performance"; though there is evidence for the construct validity of performance ratings, there is also clear evidence that factors other than performance influence ratings, regardless of the source (Murphy and Cleveland, 1995).

Multisource ratings are more likely to provide valid information about how one's performance is perceived at various levels of the organization, and regardless of whether these perceptions are accurate, it should be useful information to find out that your subordinates, peers, and so on believe that you are effective or ineffective in various aspects of your job.

Perceptions of Validity and Users' Acceptance of Rating Information

Validity can be assessed statistically, correlating between ratings with indicators of individual and organizational success (including promotion, salary, and so forth), but in evaluating the validity and utility of multisource ratings it is important to go beyond simple statistical procedures. In particular, it is important to determine whether the participants themselves believe that multisource ratings provide valid and useful information about performance. Ratings are unlikely to be useful or to lead to meaningful change unless raters and ratees accept them as valid indicators of performance and effectiveness (see Chapter Thirty).

There is evidence that perceived validity of MSF depends not only on who is being evaluated but also on which performance areas or competencies are assessed and on the purpose or use of ratings. For example, supervisors are likely to believe that subordinates can evaluate some dimensions of their job but not others (McEvoy, 1990). Dimensions or competencies that managers believe can be reasonably assessed by subordinates include leadership, oral communication, delegation, coordination, interest in subordinates, performance feedback that offers work guidance, composure and self-control, and interpersonal skills. By contrast, managers are less likely to believe subordinates can evaluate such dimensions as planning and organizing, budgeting, goal setting, decision making, creativity, quantity of work, quality of work, analytical ability, and technical ability.

There are a number of beliefs that limit supervisors' willingness to accept subordinate ratings of their performance (Bernardin, 1986). Supervisors often report that subordinates:

- Lack the information or skills needed to make valid ratings
- Are inexperienced as raters
- Have not been trained to make accurate ratings
- Harshly rate managers who are demanding
- Inflate ratings to avoid retaliation from managers
- Use ratings to undermine the authority of managers

Additionally, managers avoid organizations that use subordinate ratings, causing difficulties recruiting and retaining managers. Lastly, supervisors report subordinate ratings being nothing more than a popularity contest.

Peer ratings are often cited as a valuable source of performance feedback (Cardy and Dobbins, 1994; Wexley and Klimoski, 1984). Peers are often in a better position to evaluate job performance than supervisors, they may be more sensitive to the system factors that influence performance and how a person is able to respond to them, and they may have better understanding of the behaviors that are critical for successful job performance (Cardy and Dobbins, 1994). There is evidence that peer evaluations can be used to predict future performance, forecast final grades, and predict job advancement (Reilly and Chao, 1982; Shore, Shore, and Thornton, 1992). However, relatively little is known about the perceived validity of peer ratings.

One potential barrier to using peer ratings is the fact that employees often do not like evaluating each other; widespread dislike of rating one's peers may also interfere with the acceptance of such evaluations (Cederblom and Lounsbury, 1980; Love, 1981). Employees are more likely to accept peer ratings if appraisals are used for only developmental purposes rather than administrative (Farh, Cannella, and Bedeian, 1991). Managers are often skeptical of peer ratings because they are thought to be biased by friendship and similarity between the rater and ratee (Love, 1981) and may give

less weight to evaluations from an individual's peers than to ratings from superiors.

It is often argued that employees are more familiar with their own performance than other sources and thus are in a position to make accurate self-evaluations. Further, comparing self-ratings and supervisory ratings constitutes a method for identifying system factors that restrict performance, and for clarifying subordinate expectations about job and role requirements (Cardy and Dobbins, 1994). Although self-ratings do not show strong agreement with supervisory or peer ratings, there is evidence for the empirical validity of self-ratings in predicting objective performance. Self-ratings of performance are relatively easy to obtain and may yield at least two benefits in the feedback process: contributing to positive employee perceptions about due process and participation in important organizational decisions, and offering valuable information about system or nonindividual performance factors that have generally enhanced or inhibited effective and ineffective employee performance.

Although hardly free from bias, supervisory ratings are often accepted as valid simply because of their source. That is, the job of a manager or supervisor revolves around planning, organizing, directing, controlling, and being held accountable for accomplishing organizational objectives through his or her subordinates. Managers and supervisors are assumed to have information about the behavior and performance of their subordinates, and the task of evaluating their performance is a natural part of the supervisor's job. Multisource rating systems require out-of-role behaviors from many participants (as when subordinates are asked to evaluate their superiors); the fact that performance rating is an in-role behavior for managers and supervisors is likely to enhance the perceived validity and legitimacy of supervisory ratings.

Supervisors, peers, subordinates, and others are likely to have access to different sorts of information about an individual's performance. For example, self-raters rely more on their actual behaviors when rating, while supervisor have access to both ratee behavior and outcomes. There is evidence that supervisors rely more heavily on work outcomes than employee behaviors in making evaluations (Carson, Cardy, and Dobbins, 1991), which may help to explain why there is low agreement between supervisory and self-ratings.

Cardy and Dobbins (1994) suggest that peer raters consider both work outcomes and work behaviors when assessing performance, whereas subordinate ratings are likely be affected more by employee behavior than by outcomes. Subordinates often do not have access to the work outcomes of their supervisors and rely on direct observation of supervisory behaviors on which to based their evaluations. In general, ratings are more likely to be viewed as valid and useful by recipients if they can be confident that raters have access to the information needed to assess performance, and if they believe that raters are motivated to provide accurate evaluations. As we note below, the structure of multisource rating systems may make it easier to obtain ratings that are both accurate and credible to recipients, in comparison to ratings obtained in traditional top-down rating systems.

Is MSF Useful and Meaningful to Recipients?

The ultimate criterion for evaluating multisource feedback is probably the extent to which this information is useful and meaningful to recipients. There is an extensive literature dealing with performance feedback and its effects on individual behavior, and it is beyond the scope of this chapter to review the literature in detail. However, we can present some broad principles that are likely to affect the usefulness and meaningfulness of multisource ratings.

First, recipients' interpretation of MSF is likely to be affected by its consistency, by a comparison of information obtained from multiple perspectives with that obtained from traditional top-down evaluation systems, and by individual difference variables. Second, changes in behavior following feedback may depend on motivational factors more than on the feedback itself. Finally, perceptions of the fairness of feedback may substantially affect the success of MSF systems.

Interpreting MSF: Consistency, Value-Added, and Individual Differences

Information that is not consistent can affect interpretation of feedback, and it may also affect the degree to which behavior changes result. We have noted earlier that disagreement is common in multisource feedback. There is evidence that some individuals pay

attention to consistent information only and disregard feedback when raters disagree (Korman, 1976). These individuals may find MSF to be limited in value. However, inconsistency is not always a barrier to useful feedback. The degree of consistency can itself be a useful piece of information, because it helps determine the amount of change needed by providing comparative information to the recipient (London and Smither, 1995).

The information from MSF is likely to be evaluated in comparison to that from traditional top-down approaches, and the utility of multisource information is likely to be enhanced if recipients see added value to obtaining feedback from multiple perspectives. There is relatively little empirical research on the factors that lead recipients to perceive MSF as relatively valuable, or as redundant with traditional forms of feedback, and more work is clearly needed in this area.

From the rater's perspective, one important advantage of multisource rating is that information is often presented in an anonymous or aggregated form, whereas in traditional top-down approaches the source of ratings is known. This may give raters the ability to make more accurate ratings, and it may reduce pressures to provide the overly favorable evaluations usually encountered in top-down rating systems (Cleveland and Murphy, 1992). Enhancing the accuracy of ratings may also enhance the acceptability and perceived value of these ratings. Many of the raters in a multisource system are also ratees, and if raters believe that the structure of the rating system enhances their own ability to rate accurately, they are more likely to believe that ratings they obtain are also more accurate than would be expected under traditional top-down systems.

London and Smither (1995) identified several individual difference variables that can affect interpretation of feedback, including self-image, feedback-seeking behaviors, and self-monitoring. For example, low self-image can lead to lower self-ratings, producing an inaccurate comparison between self and other ratings. A person's propensity to concentrate on positive or negative aspects of feedback can affect his or her perceptions of the validity of the ratings received. McFarland and Miller (1994) suggest that individuals who focus on positive aspects of feedback are more likely to accept and value the feedback and are more likely to believe that they can use it to improve their own performance.

Change in Performance Following Feedback

The simple act of receiving feedback does not always change behavior. Kluger and De Nisi's 1996 meta-analysis of feedback interventions concluded that the majority of feedback recipients demonstrated a positive change in performance. However, approximately one-third of those who receive feedback showed decreases in performance. Those researchers suggest that feedback directed to the self instead of tasks is relatively ineffective in producing beneficial behavior change.

Other studies have demonstrated similar positive results in behavior and performance change following performance feedback. For example, Hazucha, Hezlett, and Schneider (1993) reported increases in managerial skill two years after receiving 360-degree feedback. Similarly, ratings of leader behaviors generally improved after feedback, (Atwater, Roush, and Fischthal, 1995). Managers whose inflated self-ratings were inconsistent with others' ratings tend to improve performance (Johnson and Ferstl, 1999). Clearly, individual perceptions of performance and behavior can change following feedback. However, change following feedback may depend on a number of factors, notably the recipient's motivation to change.

Perceptions of Fairness

Multisource assessments are thought to be more procedurally fair than traditional top-down ratings because they involve voice from each level of organization. For example, Edwards and Ewen's study (1996) of organizations that adopted MSF systems reported a 50 percent increase in perceived fairness as compared to traditional top-down systems. Multisource rating systems not only extend more opportunities for members of the organization to have some voice in evaluation but also present information in ways that reduce pressures to distort ratings. Murphy and Cleveland (1995) note that raters in a traditional system are strongly motivated to provide lenient ratings. In multisource systems, ratees often receive feedback that is anonymous or aggregated, and raters have less to fear if they give frank and accurate ratings.

Perceptions of fairness may substantially affect motivation to change behavior as a result of ratings. Raters who believe that the performance feedback they receive is unfair or biased are unlikely to expend a great deal of effort in changing their behavior; nor are they likely to accept their ratings as useful information for improving their performance in the future. Indeed, one might argue that the largest benefit of multisource rating systems is that they have the potential to deliver information in a form that is viewed as relatively fair, accurate, and unbiased. That is, the true value of these systems may lie more in the credibility of the information than in the fact that they produce more information than what is usually obtained from top-down rating systems.

Conclusions

Raters, ratees, and organizations are likely to emphasize different criteria in evaluating multisource ratings. From the perspective of the ratee, the most important issue is probably the extent to which ratings yield valid and meaningful information that helps them improve their performance and effectiveness. Reliability and validity are both important facets of this evaluation (meaning, if ratings are unreliable, they are unlikely to be useful), but the credibility and perceived fairness of ratings is also important.

From the rater's perspective, the most important criteria may be those that are tied to how ratings are collected and used (are ratings aggregated?). Organizations are likely to be concerned with a wide range of criteria, including cost, effectiveness of feedback, legal defensibility, etc. The standard psychometric criteria (reliability, validity, and so on) are likely to be important considerations to many of the participants in multisource rating systems, but these criteria are rarely sufficient for evaluating this method of collecting and disseminating performance information.

Low levels of interrater agreement provide a real challenge to the integrity and usefulness of multisource rating systems, but the effects of disagreement between raters may have various implications depending on how the system is administered (again as an example, are ratings aggregated?) and, more important, how multisource ratings are used. Performance ratings are used for a wide range of purposes in organizations (Cleveland, Murphy, and Williams, 1989), and it is likely that multisource ratings may also serve a number of purposes. Disagreements between raters may give useful information when providing developmental feedback, but they may undermine the credibility and defensibility of the same ratings if they are used for administrative purposes (salary, promotion). The relative emphasis given to different criteria (reliability, validity, perceived validity, fairness perceptions, etc.) in evaluating multisource rating systems is likely to depend on the purposes and goals of the system.

References

Atwater, L. E., Roush, P., and Fischthal, A. "The Influence of Upward Feedback on Self and Follower Ratings of Leadership." *Personnel Psychology,* 1995, *48,* 35–59.

Bernardin, H. J. "Subordinate Appraisal: A Valuable Source of Information About Managers." *Human Resource Management,* 1986, *25,* 421–439.

Bommer, W. H., and others. "On the Interchangeability of Objective and Subjective Measures of Employee Performance: A Meta-Analysis." *Personnel Psychology,* 1995, *48,* 587–605.

Borman, W. C. "The Rating of Individuals in Organizations: An Alternative Approach." *Organizational Behavior and Human Performance,* 1974, *12,* 105–124.

Bozeman, D. P. "Interrater Agreement in Multisource Performance Appraisal: A Commentary." *Journal of Organizational Behavior,* 1997, *18,* 313–316.

Bretz, R. D., Milkovich, G. T., and Read, W. "The Current State of Performance Research and Practice: Concerns, Directions, and Implications." *Journal of Management,* 1992, *18,* 321–352.

Cardy, R. L., and Dobbins, G. H. *Performance Appraisal: Alternative Perspectives.* Cincinnati, Ohio: South-Western, 1994.

Carson, K. P., Cardy, R. L., and Dobbins, G. H. "Performance Appraisal as Effective Management or Deadly Management Disease: Two Initial Empirical Investigations." *Group and Organization Studies,* 1991, *16,* 143–159.

Cederblom, D., and Lounsbury, J. W. "An Investigation of User Acceptance of Peer Evaluations." *Personnel Psychology,* 1980, *33,* 567–579.

Church, A. H., and Bracken, D. W. "Advancing the State of the Art of 360-Degree Feedback: Guest Editors' Comments on the Research and Practice of Multirater Assessment Methods." *Group and Organization Management,* 1997, *22,* 149–161.

Cleveland, J. N., and Murphy, K. R. "Analyzing Performance Appraisal as Goal-Directed Behavior." In G. Ferris and K. Rowland (eds.), *Research in Personnel and Human Resources Management.* Vol. 10. Greenwich, Conn.: JAI Press, 1992.

Cleveland, J. N., Murphy, K. R., and Williams, R. "Multiple Uses of Performance Appraisal: Prevalence and Correlates." *Journal of Applied Psychology,* 1989, *74,* 130–135.

Conway, J. M., and Huffcutt, A. I. "Psychometric Properties of Multisource Performance Ratings: A Meta-Analysis of Subordinate, Supervisor, Peer, and Self-Ratings." *Human Performance,* 1997, *10,* 331–360.

Conway, J. M., Lowe, K. L., and Langley, K. C. "Peer and Subordinate Ratings and Objective Performance, Ability, and Personality: A Meta-Analysis." Unpublished manuscript, 1999.

Edwards, M. R., and Ewen, A. J. *360-Degree Feedback: The Powerful New Model for Employee Assessment and Performance Improvement.* New York: AMACOM, 1996.

Farh, J. L., Cannella, A. A., and Bedeian, A. G. "Peer Ratings: The Impact of Purpose on Rating Quality and User Acceptance." *Group and Organization Studies,* 1991, *16,* 367–386.

Greguras, G. J., and Robie, C. "A New Look at Within-Source Interrater Reliability of 360-Degree Feedback Ratings." *Journal of Applied Psychology,* 1998, *83,* 960–968.

Harris, M. H., and Schaubroeck, J. "A Meta-Analysis of Self-Supervisory, Self-Peer, and Peer-Supervisory Ratings." *Personnel Psychology,* 1988, *41,* 43–62.

Hazucha, J. F., Hezlett, S. A., and Schneider, R. J. "The Impact of 360-Degree Feedback on Management Skills Development." *Human Resource Management,* 1993, *32,* 325–351.

Heneman, R. L. "The Relationship Between Supervisory Ratings and Results-Oriented Measures of Performance: A Meta-Analysis." *Personnel Psychology,* 1986, *39,* 811–826.

Johnson, J. W., and Ferstl, K. L. "The Effects of Interrater and Self-Other Agreement on Performance Improvement Following Upward Feedback." *Personnel Psychology,* 1999, *52,* 272–303.

Kluger, A. N., and De Nisi, A. "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory." *Psychological Bulletin,* 1996, *119,* 254–284.

Korman, A. K. "Hypothesis of Work Behavior Revisited and an Extension." *Academy of Management Review,* 1976, *1,* 50–63.

Landy, F. J., and Farr, J. L. "Performance Rating." *Psychological Bulletin,* 1980, *87,* 72–107.

London, M., and Smither, J. W. "Can Multisource Feedback Change Perceptions of Goal Accomplishment, Self-Evaluations, and Performance-Related Outcomes? Theory-Based Applications and Directions for Research." *Personnel Psychology,* 1995, *48,* 803–839.

London, M., Smither, J. W., and Adsit, D. J. "Accountability: The Achilles' Heel of Multisource Feedback." *Group and Organization Management,* 1997, *22,* 162–184.

Long, P. *Performance Appraisal Revisited.* London: Institute of Personnel Management, 1986.

Longenecker, C. O., Sims, H. P., and Gioia, D. A. "Behind the Mask: The Politics of Employee Appraisal." *Academy of Management Executive,* 1987, *1,* 183–193.

Love, K. G. "Comparison of Peer Assessment Methods: Reliability, Validity, Friendship Bias, and User Reaction." *Journal of Applied Psychology,* 1981, *66,* 451–457.

McEvoy, G. M. "Public Sector Managers' Reactions to Appraisals by Subordinates." *Public Personnel Management,* 1990, *19,* 201–212.

McFarland, C., and Miller, D. T. "The Framing of Relative Performance Feedback: Seeing the Glass as Half Empty or Half Full." *Journal of Personality and Social Psychology,* 1994, *66,* 1061–1073.

Milkovich, G. T., and Wigdor, A. K. *Pay for Performance.* Washington, D.C.: National Academy Press, 1991.

Mount, M. K., and others. "Trait, Rater and Level Effects in 360-Degree Performance Ratings." *Personnel Psychology,* 1998, *51,* 557–576.

Murphy, K., and Cleveland, J. *Understanding Performance Appraisal: Social, Organizational, and Goal-Oriented Perspectives.* Thousand Oaks, Calif.: Sage, 1995.

Reilly, R. R., and Chao, G. T. "Validity and Fairness of Some Alternate Employee Selection Procedures." *Personnel Psychology,* 1982, *35,* 1–67.

Schmidt, F. L., and Hunter, J. E. "Measurement Error in Psychological Research: Lessons from 26 Research Scenarios." *Psychological Methods,* 1996, *1,* 199–223.

Schmidt, F. L, Hunter, J. E., and Caplan, R. "Validity Generalization Results for Two Jobs in the Petroleum Industry." *Journal of Applied Psychology,* 1981, *66,* 261–273.

Scullen, S. E. "When Ratings from One Source Have Been Averaged, But Ratings from Another Source Have Not: Problems and Solutions." *Journal of Applied Psychology,* 1997, *82,* 880–888.

Scullen, S. E., Mount, M. K., and Goff, M. "Understanding the Latent Structure of Job Performance Ratings." *Journal of Applied Psychology,* forthcoming.

Shore, T. H., Shore, L. M., and Thornton, G. C. "Construct Validity of

Self- and Peer Evaluations of Performance Dimensions in an Assessment Center." *Journal of Applied Psychology*, 1992, *77*, 42–54.
Thornton, G. C., III. "Psychometric Properties of Self-Appraisals of Job Performance." *Personnel Psychology*, 1980, *33*, 263–271.
Viswesvaran, C., Ones, D. S., and Schmidt, F. L. "Comparative Analysis of the Reliability of Job Performance Ratings." *Journal of Applied Psychology*, 1996, *81*, 557–574.
Wexley, K. N., and Klimoski, R. J. "Performance Appraisal: An Update." In G. R. Ferris and K. M. Rowland (eds.), *Research in Personnel and Human Resources Management*. Vol. 2. Greenwich, Conn.: JAI Press, 1984.
Woehr, D. J., Sheehan, M. K., and Bennett, W. "Understanding Disagreement Across Rating Sources: An Assessment of the Measurement Equivalence of Raters in 360-Degree Feedback Systems." Paper presented at the fourteenth annual conference of the Society for Industrial and Organizational Psychology, Atlanta, Ga., May 1999.

CHAPTER 10

Working with a Vendor for a Successful Project

Carol W. Timmreck
Tom Wentworth

Much of this handbook is devoted to design or follow-through aspects of multisource feedback (MSF). In this chapter, we explore a very important aspect of implementing the project aspects of MSF with an external vendor. Whereas some organizations handle the technical aspects of data capture and processing internally, others seek outside help for numerous reasons, among them desiring to outsource these aspects to someone whose core business it is to process MSF; or wanting to reinforce the objectivity, anonymity, and confidentiality of the process by having it handled by a third party. It is for these situations that it becomes very important to work with a "vendor" for a successful project.

Our working definition of *vendor* is an external manager of the technical aspects of project enrollment, data collection, and data processing (rather than a consultant who offers a broader range of consulting services such as content development, or follow-up training and coaching). We address this topic from our perspectives and professional experience as an internal consultant and a processing vendor. Many views expressed in this chapter reflect a bias to outsource this process to a competent, experienced vendor; the bias comes from the experience and perspective of the internal consultant accountable to management for the quality of the outcome.