The identification and choice of criteria for evaluating the quality of performance ratings (i.e., rating effectiveness) has been recognized as a central problem for researchers and practitioners since the early work of Weitz (1961), Wallace (1965), and Smith (1976). A large number of criteria have been suggested (e.g., several rating error indices and rating accuracy indices) with no rationale for choosing among them. Recent research has been highly critical of these available criteria, but has not yet proposed a solution to the criterion problem (Guion & Gibson, 1987; Murphy & Balzer, 1989; Sulsky & Balzer, 1988; Wexley, 1987).

As we prepared this chapter, it became quite clear that, despite admonitions by Weitz (1961), Wallace (1965), and Smith (1976), we still have not come to grips with problems in evaluating and choosing performance appraisal criteria.

In the first section of this chapter, we draw three general conclusions that may not be new, but merit repetition:

1. Some of our most popular measures of performance appraisal effectiveness, namely rating error and rating accuracy indices, have serious conceptual and methodological flaws. Thus, it is questionable whether research using these measures to evaluate performance appraisal systems tells us much about performance appraisal effectiveness.

2. Most of our measures of performance appraisal effectiveness are those that are of interest to performance appraisal researchers, but not to other constituents of the performance rating process such as company executives, personnel managers, raters, and ratees.

3. Researchers and practitioners have not adequately investigated how (or whether) performance appraisal effectiveness contributes to more distal indices of organizational effectiveness, such as productivity or profitability.
In light of these problems, we will attempt to expand our conceptualization of appraisal effectiveness to encompass a broader range of effectiveness criteria, including measures that may reflect the principal interests of organizations—whether employee productivity improved and whether personnel programs are cost-effective? Furthermore, we will attempt to show that the “correct” measures of performance appraisal effectiveness differ depending on whether the individual making that determination is a researcher (who might concentrate on the pattern or distribution of performance ratings), a company executive (who might concentrate on the correctness of promotion or firing decisions), a personnel manager (who might concentrate on the acceptance and use of the appraisal system by members of the organization), an employee’s supervisor (who might concentrate on the ease of use of the rating system), or a person being evaluated by the system (who might concentrate on the accuracy and representativeness of critical incidents provided during the performance appraisal interview). Finally, given the variety of criteria that could be used to evaluate performance appraisal effectiveness, we propose a framework for formalizing the selection of criteria for evaluating appraisal systems.

**APPRAISAL CRITERIA: A BRIEF, BUT CRITICAL, REVIEW**

Many excellent papers have catalogued large numbers of measures for evaluating performance appraisal effectiveness (Bernardin & Beatty, 1984; Jacobs, Kafry, & Zedeck, 1980; Kane, 1980; Smith, 1976). But forgetting for the moment the number of criteria that are available, which criteria are we using? In preparing this chapter, we reviewed the major industrial/organizational (I/O) psychology journals (Journal of Applied Psychology, Personnel Psychology, Organizational Behavior & Human Decision Processes, and Academy of Management Journal) for the years 1976 and 1986 to identify the performance appraisal criteria reported during this period. The results of our review are shown in Table 7.1. The three most common types of criteria used in 1976 were rater error measures (54%), reliability and validity of ratings (31%), and rater/ratee attitudes (8%). In 1986, the use of rater/ratee attitudes increased (to 25%), while studies employing rater error measures and reliability and validity measures decreased (to 30% and 10%, respectively). In addition, rater accuracy measures became the third most common criteria used in performance appraisal studies. Evidence continues to mount, however, that seriously questions the use of each of these criteria.

---

1 We wish to thank Lance E. Anderson for his help in reviewing this literature.

### Rater Error Measures

Rater error measures, including halo, leniency/severity, central tendency, and similar-to-me error, are frequently used to investigate the quality of judgmental ratings of performance. A number of theoretical and methodological concerns have been raised regarding the use of rater error measures (Saal, Downey, & Lahey, 1980; Murphy & Balzer, 1989). Saal et al.’s (1980) review pointed out that the various conceptual and operational definitions for rater error measures have been inconsistent, thereby creating some confusion. For example, the different conceptual definitions of leniency/severity error (e.g., the assignment of ratings higher or lower than warranted given some external criterion; a shift in mean ratings from the scale midpoint) found in the studies reviewed are based on different underlying assumptions (i.e., the former requires “true” performance scores for comparison purposes, while the latter assumes that ratings which deviate from the scale midpoint are necessarily in error). And even where conceptual confusion was of less concern, multiple operational definitions make it difficult to compare and integrate research findings. For example, Fig. 7.1 presents seven different operational definitions of halo error, identified by Saal et al. (1980) and Fisicaro (1988), that have been used by various researchers and practitioners.

Saal et al. (1980) also provided some empirical evidence that both the choice of rater error measures and the operational definitions of those measures can lead to different conclusions regarding the evaluation of alternative rating scale formats. Finally, the authors noted that different approaches for collecting appropri-
Intercorrelations among ratee dimension ratings collapsed across raters

Factor analysis/principal component analysis of ratee dimension intercorrelation matrix

Variance of a rater’s ratings of a particular ratee across all dimension ratings

Rater x ratee interaction in a rater x ratee x dimension analysis of variance design

Median intercorrelation among ratee dimension ratings for a single rater

Sum (across dimensions and ratees) of the absolute value of illusory halo minus true halo

Sum (across dimensions and ratees) of the squared deviation of illusory halo minus true halo


### TABLE 7.2

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Halo (median correlations between performance dimensions, over ratees)</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Halo (variance of the ratings assigned to each ratee, averaged across ratees)</td>
<td>.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Leniency (absolute value of the difference between mean rating, over ratees and dimensions, and the scale midpoint)</td>
<td>-.03</td>
<td>-.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Leniency (skew of the distribution of ratings, over ratees and dimensions)</td>
<td>.15</td>
<td>-.01</td>
<td>-.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Range restriction (standard deviation of the rating distribution over ratees and dimensions)</td>
<td>-.17</td>
<td>.35</td>
<td>-.09</td>
<td>-.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Range restriction (kurtosis of the rating distribution, over ratees and dimensions)</td>
<td>-.24</td>
<td>-.02</td>
<td>.13</td>
<td>-.09</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>7. Elevation accuracy</td>
<td>-.05</td>
<td>-.02</td>
<td>-.10</td>
<td>.13</td>
<td>.02</td>
<td>.10</td>
</tr>
<tr>
<td>8. Differential elevation accuracy</td>
<td>-.06</td>
<td>.01</td>
<td>.00</td>
<td>.14</td>
<td>-.12</td>
<td>-.14</td>
</tr>
<tr>
<td>9. Stereotype accuracy</td>
<td>-.12</td>
<td>-.28</td>
<td>.00</td>
<td>-.01</td>
<td>-.07</td>
<td>-.08</td>
</tr>
<tr>
<td>10. Differential accuracy</td>
<td>-.30</td>
<td>-.50</td>
<td>.05</td>
<td>.00</td>
<td>-.10</td>
<td>-.10</td>
</tr>
</tbody>
</table>

Based on Murphy and Balzer (1989). Meta-analysis (N = 1096, analysis includes 10 studies).

A second, more pessimistic conclusion on the usefulness of rater error indices was provided by Murphy and Balzer (1989). They pointed out that rater error measures are typically used to provide indirect measures of rating accuracy when direct measures of accuracy are unavailable or difficult to obtain. But what if direct measures of accuracy are available? Are rater error measures related to direct measures of rating accuracy? Murphy and Balzer examined relationships between rater error and rater accuracy indices in ten separate data sets. Four measures of rating accuracy (elevation, differential elevation, stereotype accuracy, and differential accuracy; Cronbach, 1955) were computed for each subject. In addition, six rater error measures (two each for halo, leniency, and central tendency) discussed by Saal et al. (1980) were also computed. The average correlations among rater error and rating accuracy measures, weighting correlations from each study according to the study’s sample size, are shown in Table 7.2. Their major conclusion was that none of the rater error measures showed consistent correlations with any of the accuracy measures. Furthermore, alternative measures of halo, leniency, and range restriction showed little equivalence (this was also suggested by Saal et al., 1980). Murphy and Balzer (1989) concluded that error measures provide little information about accuracy, and may in fact reflect real group differences in the pattern and level of performance. They also concluded that the validity of rater error measures is questionable, and that great caution should be exercised in their use.

### Conclusions

Although rater error measures have been used widely by researchers to evaluate the effectiveness of different rating scale formats and rater training programs, both theory and empirical evidence casts doubt on the validity of the conclusions drawn in this research. Perhaps many of the “mixed” patterns of findings reported in the performance appraisal literature (for a review, see Landy & Farr, 1983) are due largely to the use of different rater-error criteria across studies. Although reanalysis of earlier studies’ data using consistent operational definitions of rater error measures may generate consistent results across studies, the issue of validity would remain unexamined.

### Rater Accuracy Measures

Accuracy scores are typically computed by comparing a rater’s performance evaluations for n ratees on k performance dimensions with corresponding evalu-
ons provided by “expert” raters. The closer the rater’s ratings are to the experts’ ratings, the more accurate those ratings are thought to be (Borman, 1977). The primary advantage to using accuracy scores is that they provide a direct, rather than indirect, measure of accuracy. Furthermore, whereas rater error scores are often computed under the assumptions that performance ratings are normally distributed and correlations among rating dimensions are essentially zero, accuracy scores require no assumptions about the actual distribution of ratee performance or the “true” intercorrelations among rating dimensions.

As we have seen earlier, accuracy scores are becoming more popular in performance rating research (Sulsky & Balzer, 1988; for applications of accuracy scores in research on the selection interview and job evaluation, see Dipboye, tramler, & Fontenelle, 1984; Hahn & Dipboye, 1988; Vance, Kuhnert, & Farr, 1978). In addition, halo and leniency have been reconfigured to allow computation of “illusory halo” and “illusory leniency” based on comparisons between raters’ and experts’ ratings (Bingham, 1939; Cooper, 1981; Murphy & Balzer, 1989). But rater accuracy measures are also characterized by conceptual and methodological problems. A recent paper by Sulsky and Balzer (1988) provides a critical analysis of the accuracy measures used in performance appraisal research. Four distinct criticisms are raised.

1. Current measures of rating accuracy are based on different conceptualizations of accuracy.

Sulsky and Balzer (1988) identified a variety of conceptualizations of accuracy. For example, Borman (1977) conceptualized rating accuracy as the relationship between the ratings and a second set of ratings considered to be an acceptable standard for comparison. This second set of ratings is usually gathered from expert raters and has been termed “true score” ratings. Alternatively, Cronbach (1955) conceptualized accuracy as the distance between the ratings and corresponding true score ratings. This latter conceptualization of accuracy is more consistent with other psychometric definitions of accuracy (e.g., Guion, 1965), while the former might more appropriately be labeled an indicator of rater validity.

2. Different operational definitions of accuracy exist and result in a variety of measures that do not show evidence of convergence.

To illustrate the importance of different operationalizations of accuracy measures, data from two studies were used to compute eight different accuracy scores. These accuracy scores were then correlated to assess whether their interrelationships were strong and positive. These correlations, shown in Table 7.3, show weak relationships among the various measures, with an average correlation of .19. Taken together, these findings suggest that different accuracy mea-

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Elevation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Differential</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elevation</td>
<td>.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Stereotype</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>.11</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Differential</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>.12</td>
<td>.04</td>
<td>.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Norman’s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differential</td>
<td>.18</td>
<td>.41</td>
<td>.16</td>
<td>.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Halo Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>1.00</td>
<td>-.04</td>
<td>.09</td>
<td>.12</td>
<td>.18</td>
<td>.66</td>
<td>.67</td>
<td></td>
</tr>
</tbody>
</table>

Correlations not computed because relationships are data dependent.
Adapted from Sulsky and Balzer (1988).

3. Procedures for estimating true scores have been inconsistent.

Despite the importance of true scores as a standard for comparison, procedures for estimating true scores vary considerably, ranging from simply averaging ratings from undergraduate raters (Bernard & Pence, 1980) to elaborately training experts and formally examining their ratings for intrarater reliability and discrimination among raters and rating dimensions (Borman, 1977). Clearly, averaging scores from nonexpert raters should not imply agreement among raters on the true level of ratee performance, because the dispersion around this average score can be quite large. Also, asking experts who disagree to reach consensus on a final score may result in artificial agreement. In sum, the quality of the resulting estimates may be highly questionable, notwithstanding the use of experts.

4. Accuracy scores have limited usefulness.

Although accuracy may be an important criterion for evaluating the effectiveness of performance appraisal ratings, it is by no means the ultimate criterion. Sulsky and Balzer (1988) argued that accuracy should not be the only perfor-
performance appraisal criterion of interest to researchers and practitioners. Validity and reliability, as discussed in the next section, can be useful criteria when decisions require only that raters correctly rank order rates for organizational decisions (e.g., promotions, deciding who should receive first access to training, etc.). Alternatively, accuracy may be important for decisions based on a cutoff score (requiring correspondence with some "real-world" state of affairs). In other words, we need to establish the purposes or objectives for performance appraisal, and choose criteria that will tell us whether we are meeting those objectives.

Conclusions. Despite the growing popularity of accuracy scores, theoretical and methodological limitations are as serious for these measures as for rater error scores. Adoption of accuracy measures as substitutes for error measures only replaces one set of theoretical and methodological concerns with another. Hence, while useful, accuracy scores must be viewed in light of their potential limitations, and as only one of many criteria that may be used to evaluate the effectiveness of performance ratings.

Reliability and Validity Measures

Two concerns are noted regarding the use of reliability and validity of performance ratings as criteria for rating effectiveness: (a) they are underused, and (b) they have been confused with accuracy. Our own impression (Sulsky & Balzer, 1988) is that in the excitement over the prospect of directly computing performance raters' accuracy, researchers neglected reliability and validity criteria. Perhaps it was because reliability and validity were necessary but not sufficient conditions for accuracy (i.e., they provide information on the strength of the relationship between a set of scores and a corresponding set of true scores, whereas accuracy also includes information about the nature of the relationship between the two score distributions; Gordon, 1970). Thus, accuracy scores seemed to be more important. Or perhaps it was because the relationship between reliability, validity, and accuracy is unclear. Borman's (1977) measure of differential accuracy, for example, is more precisely (using Gordon's, 1970, and Guion's, 1965, definition of accuracy) a measure of rating validity.

For whatever reason, reliability and validity have been less prominent measures of rating effectiveness than error and accuracy measures. This is unfortunate because the information provided by reliability and validity can be important effectiveness criteria. When consistent and correct rank ordering of rates is important for the purpose at hand, reliability and validity should be the criteria of choice.

As a final point, it is worth noting that several authors (e.g., James, 1973; Smith, 1976) have called for research on the construct validity of ratings. Unfortunately, research in this vein has been rather difficult to do, in part because the "performance" construct is still not well-defined (Campbell, 1983; Guion, 1983; Hunter, 1983; Naylor, 1983). Nevertheless, some interesting work has been completed in this area (Nathan & Alexander, 1988; Pulakos, Borman, & Hough, 1988; Vance, MacCallum, Coover, & Hedge, 1988).

Rater/Ratee Attitudes and Perceptions

A number of rater/ratee attitudes toward, and perceptions of, performance appraisal have been investigated, including satisfaction with appraisal ratings, perceptions of fairness, feedback, and appraisal interviews, rating format ease of use and objectivity, meaningfulness and accuracy of feedback, and perceptions of appraisal system operation and organizational impact (Bannister, 1986; Barr, Brief, & Fuk, 1981; Dipboye & DePoutrriard, 1981; Dorfman, Steven, & Loveland, 1986; Landy, Barnes, & Murphy, 1978; Pearce & Porter, 1986; Wiersma & Latham, 1986). For example, Landy et al. (1978) found that appraisal system characteristics (e.g., the frequency of performance evaluations, the opportunity for raters to express their opinions during the evaluation process) influenced perceptions of fairness in performance evaluations. Dipboye and DePoutrriard (1981) found that employee satisfaction with the appraisal process was related to employee participation during the performance feedback process, to discussion of plans or objectives for future performance, and to perceptions that evaluations were based on relevant work factors.

Many of the attitudinal scales used in this research appear to be well constructed, and information on scale development and psychometric properties (predominantly internal consistency reliability) is available. A number of attitudinal measures are questionable, however, some use single-item scales with no evidence of reliability or validity, some adapt previously existing scales (and thereby potentially affect their psychometric properties), and some fail to provide any evidence of scale validity. Thus, a careful methodological analysis of reported rater/ratee attitudinal measures, particularly their validity, should be conducted prior to accepting a study's conclusions.

Summary

No performance appraisal criterion measure is perfect. Each can, however, provide useful information given an adequate understanding of the conceptual and methodological issues underlying the measure. Most important, careful consideration should be given to the choice of performance appraisal criteria. Criteria must be chosen that address the issue (or issues) one is interested in, whether it be rating distributions, correct rank-ordering of performance, or distance from some true level of performance.
of performance appraisal. For example, Banks and Murphy noted that current investigations of raters' cognitive processes may not be useful to practitioners in terms of suggesting ways to improve performance appraisal procedures. Although Banks and Murphy focused primarily on the independent variables in performance appraisal research, we believe the argument should also be extended to the dependent variables used. Specifically, we argue in the remainder of this chapter that the dependent variables (or criteria) traditionally measured by performance appraisal researchers may be of little or no interest to practitioners. Alternatively, certain criteria of performance appraisal effectiveness that might be of interest to practitioners have been either largely or completely ignored in performance appraisal research.

Perhaps at the core of this apparent "criterion gap" is a failure to appreciate the inherent complexity of the meaning of performance appraisal effectiveness. Researchers appear to have equated performance appraisal effectiveness with the psychometric quality of ratings; thus, error-free ratings are effective ratings, and effective ratings imply an effective performance appraisal system. Practitioners, however, may be concerned with other criteria of performance appraisal effectiveness (e.g., Do the employees accept their ratings? Do ratings influence future levels of performance?). Ultimately, however, the concept of performance appraisal effectiveness is perhaps best characterized as a larger, multidimensional construct; it will include a variety of specific criteria (e.g., rater error indices, employee attitudes toward appraisal, etc.) that collectively comprise the construct of performance appraisal effectiveness.

If we accept the premise that a researcher-practitioner gap exists in terms of measuring performance appraisal effectiveness, this will diminish the useful contributions of science to practice and practice to science in performance appraisal. Published performance appraisal research may be viewed by practitioners as an academic exercise of limited value to organizations struggling with the task of appraisal (e.g., Dossett, Feldman, Timmreck, & Vandaveer, 1989), and practitioners' contributions to the area may be ignored by researchers (perhaps due to the lack of experimental controls in field research), or more seriously, not submitted to professional meetings or scholarly journals (Guion, 1988).

Although the source(s) of this gap may be difficult to trace, we believe part of the difficulty can be explained by the fact that researchers have not shown enough concern for the process of choosing criteria used to define performance appraisal. In a laboratory situation where a researcher is investigating raters' conceptual similarity error, the choice of a halo measure as a criterion is clearly appropriate. In some laboratory and field research, however, it is less clear what rating criterion (or criteria) should be used. For example, when evaluating a new rater training program, what properties should be included as criteria: amount of rater halo, accuracy of ratings, satisfaction with training program, and so forth? Components of the performance appraisal system can be evaluated on a number of conceptually and methodologically different criteria, but we have no clear procedure for choosing which criteria to use in specific situations.

We believe that researchers too often choose criteria on the basis of inappropriate considerations, including: (a) they are familiar criteria that have been used in the past; (b) they are newly discovered criteria that are thought to be the wave of the future; or (c) they are readily available criteria that are easy and inexpensive to gather. Weitz (1961) made a similar point when he pointed out that criteria are not chosen because of some careful understanding of the construct "performance," but rather are chosen because the criteria are thought to be relevant, have been used by other investigators, or are readily available (see also Guion, 1961). Thus, we see laboratory studies with videotapes of ratee behavior that use rater accuracy scores (given their relative ease of computation and their use in other lab studies), and field studies using rater error measures (because estimating true scores in the field is extremely difficult). Although these choices are understandable, these are not sufficient reasons for choosing criteria. Criteria may seem sensible and interesting to the researcher, but they may be of no interest or relevance to other constituents of the performance appraisal system (Strasser & Bateman, 1984). Clearly, both scientists and practitioners need a more systematic rationale for choosing criteria to evaluate the effectiveness of performance appraisal systems.

Performance Appraisal Criteria: Constituencies and Goals

In choosing criteria to evaluate a performance appraisal system, it is important to consider the goal(s) and purpose(s) of appraisal in the organization. We believe that any framework for choosing criteria should emphasize that the choice of criteria depends on the goals of the performance appraisal system; others have made similar suggestions in the past (e.g., Smith, 1976; Sulsky & Balzer, 1988; Weitz, 1961). Furthermore, the framework should recognize that different individuals or constituencies involved in the performance appraisal system may have similar, overlapping, or different goals when evaluating the effectiveness of a system. Different goals may lead to different choices of what aspects of the performance appraisal system should be examined, and which criteria should be chosen to examine those aspects.

As an example, four different constituencies can be identified who have some interest in evaluating the effectiveness of a performance appraisal system: (a) the organization that sponsors and supports the performance appraisal system; (b) raters responsible for conducting performance appraisals; (c) ratees who are evaluated by the appraisal system; and (d) researchers internal (e.g., individuals from the company's human resource department) or external (e.g., academicians) to the organization. Other constituencies may also be identified (e.g., personnel technicians responsible for administering and compiling performance appraisal information), but these four interest groups represent most of the individuals with vested interests in the performance appraisal system. Within each constituent interest group, different individuals or subgroups may also have
different interests and goals when evaluating the performance appraisal system. For example, raters responsible for evaluating many rates may have different notions of the effectiveness of the performance appraisal system than do raters who are required to evaluate only one ratee.

Each constituent interest group can have a number of goals for the performance appraisal system. For example, an organization's goals for the performance appraisal system may include increased employee productivity, higher company profits, defensibility in court against charges of discrimination, and compatibility with other organizational functions (e.g., compensation or production). They may also have different goals for the performance appraisal system. Raters may focus on how easy the system is to use, on positive ratee attitudes toward the performance appraisal system, on improvements in ratee work motivation, and/or on the absence of role conflict or role overload for the raters (e.g., the performance appraisal system's workload does not interfere with the performance of other responsibilities). Ratee goals may include a performance appraisal system that can be trusted, that clarifies instrumentalities between job performance and organizational rewards, that provides fair access to rewards, and that provides accurate and detailed feedback that can be used to direct future performance. Finally, researchers' goals may include ratings that are highly accurate and free from traditional rater errors, that allow for evaluation of performance appraisal system components (e.g., rating format, rater training, etc.), that lead to an understanding of raters' underlying psychological processes, and rater/ratee attitudes toward performance appraisal and their relationship to other work attitudes (e.g., job satisfaction) and behaviors (e.g., turnover).

In summary, a wide variety of goals can be identified for different groups of individuals interested in evaluating performance appraisal systems. Two additional points can also be made. First, different individuals within a constituent interest group may have unique performance appraisal system goals. For example, some researchers may be interested solely in the psychometric qualities of ratings, whereas other researchers may be interested in information processing capabilities (recallability of performance appraisal information, strategies for integrating appraisal information). Second, it is possible that the goals of different constituencies may overlap. For example, both rater and ratee constituencies may require an effective performance appraisal system to provide accurate feedback information during the performance appraisal interview. In general, however, it appears much more likely that each constituency will be interested in a narrower range of goals that are particularly relevant for that group.

Constituencies and goals: Choosing criteria. Specific categories of goals for evaluating performance appraisal effectiveness are proposed to direct attention toward certain features of the performance appraisal system. The performance appraisal system can be broken down into three components: (a) inputs to the performance appraisal system, the components in place prior to the evaluation of performance (e.g., rating purpose, rater selection, rating instrument, rater training, etc.); (b) throughputs of the performance appraisal process, the components involved in the appraisal of performance (e.g., feedback, performance monitoring, completing the rating instrument, etc.); and (c) outputs of the performance appraisal system, the components that result from the appraisal of performance (e.g., performance ratings, personnel decisions based on these ratings, characteristics of feedback, etc.). These aspects of the performance appraisal system provide sources of information for evaluating performance appraisal effectiveness.

The goals of a constituent of the performance appraisal system should direct the investigator's attention toward particular outcomes or products of the performance appraisal system, and this should in turn influence the choice of criteria. As shown in Fig. 7.2, if the goal is compatibility with other systems, a number of components of the appraisal system may be of interest: rating purpose (i.e., is performance appraisal providing information necessary for compensation, planning, etc.), selecting raters, rating schedule, performance monitoring (i.e., is performance appraisal interfering with rater/ratee abilities to fulfill other organizational roles), and goals (i.e., are performance appraisal goals contributing to all required areas of job performance). If the goal is the rater's ease of use, an investigator might instead investigate only input and throughput characteristics (e.g., How much training is required? How much time will performance monitoring take? Is the rating instrument easy to complete? Does the feedback meeting take too much time, and is it difficult to conduct?).

Our model for selecting criteria for evaluating the effectiveness of performance appraisal systems proposes that effectiveness is determined by the goals of particular constituents of the performance appraisal system, and that these goals may lead investigators to focus on a subset of information regarding the performance appraisal system. It is this information that should determine the actual criteria chosen. The use of this framework leads to a number of conclusions regarding research on performance appraisal effectiveness.

First, as noted earlier, performance appraisal effectiveness should be viewed as a multidimensional construct. Multiple constituencies and goals imply that a single criterion of effectiveness will be deficient for measuring the complex construct of performance appraisal effectiveness. Thus, multiple criteria would appear to be critical for evaluating performance appraisal effectiveness.

Second, the selection of criteria for evaluating performance appraisal effectiveness should recognize that "effectiveness" means different things to different people. Different groups of individuals involved in the appraisal process have particular and unique goals for evaluating effectiveness. Thus, effectiveness criteria should be carefully reviewed to determine whether they reflect the goals of all constituents of the performance appraisal system.

Third, our review suggests that past research has focused on a limited number of criteria such as the psychometric qualities of ratings, rater/ratee attitudes, and
7. PERFORMANCE APPRAISAL EFFECTIVENESS

Given that researchers and practitioners have often neglected to consider criteria that might be of interest to other constituencies, what can we do to increase their awareness to a broader set of criteria for effectiveness? One solution is a comprehensive framework to guide the choice of effectiveness criteria; such a framework would serve to close the existing gap between scientists, practitioners, and other constituencies. Here, we propose such a framework and describe how it can help us identify and choose criteria of interest to one constituency—the organization. We focus our example on how one might identify organizationally relevant criteria that can be used to evaluate a performance appraisal system. Our decision to use the organization as the constituency was not an arbitrary one; published research that addresses the impact of performance appraisal systems on organizational goals such as productivity and profit, goals that have been seen as “the ultimate effectiveness criteria” for many organizations (Katz & Kahn, 1978), is relatively scarce.

Clearly, organizations may have interests that overlap with those of other constituents of the performance appraisal process. All constituent groups may have an interest, for example, in the attitudes expressed toward the performance appraisal process or the correctness of the rankings of individuals in line for promotion. But what about those goals that may be more specific to the organization such as increased performance and productivity? To be sure, reviews of research on appraisal feedback (Guzzo & Bondy, 1983; Kopelman, 1986) and goal setting (Guzzo & Bondy, 1983; Latham & Baltes, 1975) suggest that characteristics of the performance appraisal system do affect “bottom-line” measures of performance (i.e., the quantity and quality of performance) and productivity; recent applications of utility measurement (Casio, 1987) may enhance our ability to estimate dollar savings to an organization due to certain aspects (e.g., feedback, rater training, etc.) of the performance appraisal system (Landy, Farr, & Jacobs, 1982). But as Smith (1976) pointed out, measuring the impact of appraisal systems on organizational goals remains elusive because these criteria

---

**FIG. 7.2.** A framework with examples for selecting criteria for performance appraisal effectiveness.
do not easily lend themselves to measurement, because the financial costs associated with measuring them are often prohibitive, and because of the distal relationship of these criteria to the characteristics of the appraisal system. Difficulties inherent in conducting rigorous and internally valid studies using such criteria may also be responsible for the de-emphasis on organizational goals and the emphasis instead on researcher goals by methodologically rigorous researchers. In any event, given real world constraints, measuring organizational goals may prove to be extremely difficult.

Organizational Objectives: Measures for Realizing Organizational Goals

It is very difficult to determine whether an appraisal system is influencing the attainment of organizational goals such as profit and productivity. It may, however, be easier to evaluate whether the appraisal system is meeting organizational objectives. Our use of the term is not simply semantic. Objectives differ from goals in terms of scope, specificity, and proximity to performance appraisal system characteristics. Thus, an objective is more narrow than a goal, and suggests a specific measure of the goal of interest. For example, if the organizational goal is productivity, specific objectives that must be reached to obtain this goal include: (a) promote the best worker (who should be the most productive); (b) provide valid appraisal feedback to rates (thereby enhancing work performance); and (c) develop performance goals (because goals have been found to lead to increased performance, e.g., Latham & Kinne, 1974). If the organizational goal is an appraisal system that permits fair access of minorities to upper-level positions, objectives may be: (a) rating distributions that yield similar distributions for minority and majority employees; (b) personnel decisions that promote a certain proportion of minorities; or (c) that minority rates who perceive the appraisal system as fair.

It should be noted that achieving objectives does not guarantee goal attainment, although achieving objectives should help in this regard. Thus, forced promotion of a proportion of minorities before some are adequately prepared may lead to voluntary and involuntary turnover that inhibits goal achievement. Similarly, even if we promote the most qualified workers, productivity may suffer due to a host of factors unrelated to worker qualifications (e.g., downturn in the economy, aging equipment, etc; Landy & Farr, 1983). But even if we cannot be certain whether, or to what extent, the appraisal system is having a positive impact upon distal organizational goals, evaluating objectives related to these goals provides useful information regarding the quality of the system outputs. That is, even if we cannot guarantee that correct personnel decisions lead to increased performance, we are in a position to evaluate whether correct decisions were made. If we cannot guarantee that workers are using feedback to improve performance, we can at least evaluate the quality of the feedback given.

Changing the focus from organizational goals to specific objectives related to those goals has implications for the choice of criteria. If organizational objectives are considered important, criteria should be chosen based on careful consideration of the degree to which they correspond to those organizational objectives.

An example may help clarify our point. Consider a situation where an appraisal system is being used to make promotion decisions. Two possible organizational goals may be to: (a) promote the best employees to improve productivity and, given affirmative action goals, (b) increase the proportion of minority promotions (perhaps giving little weight to differences in minority and majority performance ratings). One objective for reaching the goal of increased productivity is to promote only the most qualified individuals in the work force. This would prescribe that one criterion for evaluating the appraisal decision should focus on whether, in fact, the "best" performers (based on previous performance) were selected. A number of psychometrically oriented criterion measures could be used to examine the ratings used for promotional decisions. The choice of criterion, however, will depend on the particular decision to be made (i.e., the specific objective). If the rating system is used to promote the top n employees, our criterion need only examine the degree to which employees were correctly ranked in terms of performance levels. On the other hand, if all qualified rates (based on some cut-off performance score) are to be promoted, rating accuracy is the more important criterion, because accuracy measures evaluate the degree of correspondence between rates' performance rating levels and true levels of performance.

Focusing on the organization's goal to improve minority representation in management, one objective might be the promotion of a certain proportion of minorities into upper level positions. In this case, the criterion of interest regarding the appraisal decision is not the validity or accuracy of ratings, but rather some fairness criterion (Hunter & Schmidt, 1976). If employees' perceptions of the fairness of the personnel decisions is the organization's objective, the criterion would most likely be some attitudinal measure.

A Final Note: Identification of Organizational Goals and Objectives

The use of organizational goals and objectives as a framework for selecting criterion variables implies some agreement on what those goals and objectives are. Unfortunately, identifying agreed upon organizational goals and objectives is extremely difficult (Cameron & Whetten, 1983; Landy & Farr, 1983; Smith, 1976). As Katz and Kahn (1978) pointed out, decisions made in defining performance in organizations are not made rationally, but rather reflect bargaining and negotiation on the part of representatives of various organizational sectors or constituencies. Schein (1980) acknowledged that it may not be possible to satisfy the objectives of one constituency without adversely affecting those of another.
Thus, we are left with the basic, yet potentially impossible, challenge of specifying organizational goals and objectives in order to choose appropriate criteria. Without a clear specification of goals and objectives, the contribution of the appraisal system to the organization (if any) will remain unclear. Although we cannot adequately deal with this issue in the present chapter, we can make several recommendations.

One method for identifying what the organization considers important is to ask those organizational members responsible for developing organizational policy and strategy. One can presume that such a system-wide perspective would help prevent subgoals from particular sectors of the organization from replacing organizational goals. Or perhaps the goals of the direct users of the performance appraisal system should take precedence, particularly if the outcomes of the performance appraisal system have only distal relationships to organizational goals such as market share or profitability. In contrast, perhaps allowing key individuals who represent various areas of the organization develop organizational goals and objectives would be preferred; multiple perspectives on work processes and organizational constraints may lead to more realistic goals and objectives.

As Cameron and Whetten (1981) warned, organizational goals and objectives: (a) are often difficult to recognize or articulate, (b) differ widely among individual organizational members, and (c) change over time. Points (a) and (b) may be resolved by using judgment analytic techniques such as cognitive feedback for reducing conflicting values (e.g., Balke, Hammond, & Meyer, 1973; Balzer, Doherty, & O’Connor, 1989; Hammond & Adelman, 1976; see also Baron’s chapter in this volume). Point (c) may pose a more difficult obstacle to establishing goals and objectives, but strategies to monitor or predict changes in goals and objectives as organizations mature through their life cycles may be promising (Cameron & Whetten, 1981; Quinn & Cameron, 1983). Once goals and their accompanying objectives are defined, we can begin to identify and choose specific criterion measures for evaluating the appraisal system based on those objectives.

### AN APPLICATION OF THE FRAMEWORK FOR SELECTING CRITERIA

The proposed framework for selecting criteria based on considerations of organizational goals and objectives can be extended to other constituents of the performance appraisal process—raters, raters, or researchers. Because each of these “interest groups” may have their own goals and more specific objectives, a framework for selecting criterion measures should be useful when conceptualizing performance appraisal studies. Such a framework, including several distinct and formal steps for selecting the appropriate criteria for evaluating performance appraisal system effectiveness, is illustrated in Fig. 7.3.

<table>
<thead>
<tr>
<th>Example 1:</th>
<th>Example 2:</th>
<th>Example 3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization</td>
<td>Organization</td>
<td>Researcher</td>
</tr>
<tr>
<td>Profit</td>
<td>Profit</td>
<td>Improvements in Psychometric Quality of Promotion Ratings</td>
</tr>
<tr>
<td>Promote &quot;Best&quot; Rate from Pool of Candidates</td>
<td>Promote All Candidates Exceeding Cut-off Score</td>
<td>Reduce Leniency of Performance Ratings</td>
</tr>
<tr>
<td>Validity of Rating</td>
<td>Rater Accuracy Measure</td>
<td>Leniency Error Measure</td>
</tr>
</tbody>
</table>

**FIG. 7.3.** A framework with examples for choosing performance appraisal criteria.

### Identification of constituent group(s).

As we have suggested, an appraisal system’s effectiveness can be evaluated by a number of different constituent groups. Although each of these groups may share common goals for the appraisal process, some goals are likely to be unique to each group. If researchers ignore this important step, they may inadvertently choose criterion measures that are of peripheral value or of no interest to the practitioner, organization, or other constituencies. This is one factor that can contribute to the present researcher/practitioner gap in performance appraisal research (Banks & Murphy, 1985). Clearly identifying the constituent group will help to identify the specific goal(s) for evaluating appraisal system effectiveness (Edwards & Newman, 1982). Brief guidelines for identifying constituent groups and selecting participants from these groups were suggested by Delbecq, Van de Ven, & Gustafson (1975).

### Identification of constituent goal(s).

As discussed earlier, each constituent group can have a number of different goals. Each goal can lead to different decisions regarding how to evaluate appraisal effectiveness, which implies that it is important to clarify which goal will be the focus of the study. The decision-making literature (e.g., Delbecq et al., 1975; Edwards & Newman, 1982; Ham-

---

3Edwards and Newman (1982) distinguished between two general categories of constituents: actors, individuals who make decisions about programs and have the ability to modify programs; and stakeholders, individuals who are either directly or indirectly affected by a program. In the present context, this distinction between actors (e.g., researchers, upper-level management, personnel officers) and stakeholders (e.g., raters and raters) may lead to hypotheses about the subsequent goals and objectives of these constituent groups.
Identification of specific objective(s). Given the distal nature and vagueness of many goals, specific objectives associated with the constituent group's goal(s) should be clearly identified. This step provides the investigator with one (of many possible) specific measure of the goal of interest that can be used to evaluate the appraisal system's effectiveness.

Selection of criterion. Finally, available criteria should be reviewed, and the criterion measure(s) that assess(es) whether objectives have been met should be chosen and used to evaluate the appraisal system effectiveness.

Figure 7.3 presents three examples in which this framework is used to choose an appropriate criterion for evaluating the effectiveness of a performance appraisal system. It can be seen that differences in constituent groups, goals, and objectives all affect the characteristics required of a criterion, thereby helping us to identify the most relevant criterion for evaluating performance appraisal effectiveness. Of course, the model simplifies a very complex process. There are often many constituencies to satisfy, and multiple goals and objectives to meet. But the framework provides a strategy for allowing identifiable goals and objectives, rather than our personal biases or convictions, to guide our choice of criteria for judging the effectiveness of performance appraisal systems.

CONCLUSIONS

In this chapter, we have explored criterion measures currently used in performance appraisal research, and have highlighted some of the methodological and theoretical limitations associated with these measures. Our primary purpose, however, has been to argue that a "criterion gap" exists in performance appraisal research, and to help close this gap by providing a framework for the informed choice of criterion measures. To that end, we conceptualized performance appraisal as a process of inputs, throughputs, and outputs. We highlighted the fact that we might be interested in any of a number of different components of the overall process. Furthermore, we pointed out that a number of different constituencies or interest groups may have partially overlapping as well as distinctly different interests. Because constituents may have different interests, it is not surprising that they may choose different criterion measures for evaluating effectiveness.

We believe that the lack of communication between researchers and practitioners, and the lack of communication of both these groups with organizational members affected by the performance appraisal system, has contributed to "gaps" separating these different individuals; however, these gaps may partially reflect the fact that these individuals simply have different interests. It would appear, then, that establishing or improving communication among these groups of individuals may help close the gaps. Raters and ratees must be surveyed, interviewed, and so on to determine their views on the characteristics of an effective performance appraisal system, and these concerns should be seen as no less important than those identified by researchers and practitioners. Practitioners need to communicate with researchers, sharing problems and concerns they face when attempting to implement performance appraisal systems, and providing constructive feedback on the perceived applicability of performance appraisal research programs. This is one way in which practitioners can make an important contribution to science. Researchers, of course, must also share advances in theory, method, and so on with practitioners and other organizational members; perhaps this could be accomplished by providing informal "continuing education" for practitioners. Finally, if researchers can be more attentive to the concerns of organizations, new and previously unresearched criterion measures may evolve and become a focus of interest, and studies may be designed to examine the effects of various appraisal system-factors on these criteria, thereby providing important contributions to practice (Dosset et al., 1989).

This latter point raises an important issue that we have thus far ignored: the process of choosing criterion measures may be useful for generating hypotheses and designs for specific studies. That is, rather than designing a study and then searching for criteria to test hypotheses, the choice of criteria could actually guide the researcher during hypothesis and research design generation. Thus, for example, if satisfaction with performance feedback is deemed to be an important criterion, we might attempt to develop and test models that explain and predict satisfaction with appraisal feedback. Criterion measures may help us to ask the right questions, and may be more than simply vehicles used to provide solutions.

ACKNOWLEDGMENTS

We thank Kevin Murphy and Skip Saal for their helpful comments and criticisms on an earlier draft of this paper.

REFERENCES


Job Performance and Productivity

Kevin R. Murphy
Colorado State University

Productivity has become a national concern, and for good reason. Productivity growth in this country has slowed considerably since the late 1970s (Tuttle, 1983); in 1984, productivity grew nearly three times as quickly in Japan as in the United States. Low productivity will eventually affect our standard of living, and could lead to substantial hardships for future generations (Perloff & Nelson, 1983). The possibility that psychology can be applied to increase productivity in organizations has therefore become an increasingly important concern; productivity-oriented research has, as a result, become increasingly common in the social and behavioral sciences (Brief, 1984; Campbell & Campbell, 1988a; Kopelman, 1986).

Prior to 1980, psychologists and managers were generally pessimistic about the impact of psychologically based interventions, ranging from ability testing to goal setting, on productivity. Although researchers were aware of scattered studies demonstrating the benefits of different interventions, there was widespread doubt, both among practitioners and organizations, that these interventions were economically worthwhile. Recent research has suggested that the impact of psychologically based interventions is in fact substantial. For example, Schmidt, Hunter, McKenzie, and Muldrow (1979) estimated that the federal government could save $376 million over 10 years by using cognitive ability tests to select computer programmers. Hunter and Hunter (1984) estimated that the use of ability tests in selection for federal entry-level jobs could lead to a productivity gain of over $15 billion per year. Finally, Hunter and Schmidt (1982) estimated that the nationwide productivity gain associated with the use of valid tests in personnel selection could exceed $80 billion per year. Although Murphy (1986)