The purpose of this chapter is twofold. First, we have assessed the congruence between laboratory and field studies in three areas of research in performance appraisal (namely, rating formats, rater training, and rating purpose). Second, we have critiqued both types of research and suggested research needs and methodologies for enhancing the generalizability of results. Based on a survey of people directly involved with appraisal systems as raters, rates, or administrators, critical areas for research are identified and a modal criterion model is proposed. It is our thesis that most laboratory and field studies on performance appraisal ignore critical boundary variables (Fromkin & Streufert, 1976), lack psychological realism, and ignore major issues related to generalizability. These strong statements assume that applied researchers are concerned about the appraisal of actual job performance or what Naylor, Pritchard, and Ilgen (1980) refer to as organizational behavior. Generalizability to other appraisal situations involving performance in assessment centers, interviews, and work samples is beyond the scope of this chapter.

Fromkin and Streufert (1976) have discussed the importance of identifying boundary variables which can limit the external validity of research findings. Boundary variables are critical differences between the research setting and criterion setting which could alter the existence or magnitude of the relationship between two or more variables. For example, several studies have found a positive/negative relationship between some type of rater training program and some measure of effectiveness (for example, accuracy, halo, leniency, discriminability). Virtually every study on the effects of rater training, however, has ignored the purpose for which the data are to serve (for example, merit pay, promotion, reductions in force, feedback). Data in rater training studies are almost always collected for experimental purposes only. Thus, it may be that the significant results obtained for training in the experimental context have no generalizability to most criterion settings where the appraisal data serve some critical purpose, that is, rater motivation in these situations is unaffected by the training procedures under study.

Purpose of appraisal could also be a critical boundary variable in rating format studies. Purpose of appraisal is rarely considered in this context despite the
fact that two rating formats (forced choice and mixed standard) were specifically developed to inhibit deliberate rating distortion. Of course, the extent to which deliberate rating distortion is a problem is related to the purpose for which the data are to serve. It is no surprise that format has had little effect in rating situations where data serve no administrative purpose and the extent of rating distortion is probably minimal (Bernardin & Beatty, 1984).

Personal contact between rater and ratee (that is, life goes on after the performance appraisal) could be another boundary variable. For example, a training program found to inhibit leniency in a research setting may have no effect in a criterion setting where trained raters are also supervisors of the persons who are rated. Related to this issue is the confidentiality of ratings. Perhaps confidentiality is a boundary variable for several relationships.

In this chapter we present evidence from three sources: (1) a review of the contemporary empirical literature on performance appraisal, (2) a survey completed by personnel practitioners, and (3) another survey completed by practitioners, supervisors, and subordinates. We first conducted a review of the literature on performance appraisal (PA) research, coding studies with regard to characteristics that are pivotal in assessing their generalizability. We also conducted a survey of practitioners for purposes of developing a modal appraisal situation. Finally, we took the information from these two sources (literature and survey) and compared the characteristics of each to assess the extent of congruence between them. In so doing, we made the assumption that research with characteristics more proximal to the modal appraisal setting is more immediately generalizable to it (Cook & Campbell, 1979). For our purposes, the characteristics of the setting include any component of the appraisal process. The second survey was used to generate critical research needs from the perspective of those directly involved with ongoing appraisal systems.

**Literature Review**


Each study was coded for the following characteristics: (1) the nature of the study (lab, field, or student ratings of professor), (2) subject characteristics, (3) type of organization, (4) independent variables under study, (5) dependent variables, (6) purpose of appraisal, (7) confidentiality of ratings, (8) rate characteristics, and (9) rating format(s) under study. Due to the number of studies involving
student ratings of instructors and the unique nature of this rating situation, we have distinguished these results from results of other field settings.

A total of ninety-four studies were independently coded by two researchers, with complete agreement on all codings. The percentage of studies falling into the laboratory, field, and student rating categories were 53.2, 34, and 12.8 percent, respectively.

With regard to the independent variables under study, rating format studies were the most popular (23 percent of the research reviewed). Other frequent research topics were rater training (23 percent), rater ability factors (16 percent), and purposes of appraisal (8.5 percent).

The most frequently studied dependent variables were severity/leniency (30 percent), halo (35 percent), accuracy (19 percent), ratee discriminability (21 percent), interrater reliability (15 percent), and validity (convergent, discriminant, 11 percent).

Because of the greater frequency of research directed at the effects of rating formats and rater training, we decided to concentrate our discussion on these areas and, in addition, to review the studies on appraisal purpose. After narrowing the focus of the chapter to these three topics, we expanded our literature review in order to identify as many studies in these areas as possible.

Rating Format Comparisons

There have been numerous reviews and discussions of the results of rating format comparisons (for example, Bernardin & Beatty, 1984; Borman, in press; Green, Bernardin & Abbott, in press; Jacobs, in press). We were able to locate forty-two empirical comparisons between two or more rating formats. Twelve were lab studies using either videotaped ratees or paper-people vignettes; thirteen involved student ratings of professors; and seventeen were field studies involving police officers, computer programmers, technical engineers, department store managers, military personnel, state troopers, residence hall counselors, and several other occupations.

There were over twenty-five different dependent variables used in these studies (Bernardin & Beatty, 1984, p. 190). Regardless of the research setting or the dependent variable under study (for example, halo, leniency, discriminability, accuracy), rating format had little effect. Schmidt (1977) has argued that the most useful criterion of effectiveness across all data purposes is the disattenuated correlation between sets of ratee ratings from two formats. Most of the studies did not report the information necessary to calculate this corrected correlation. However, for the ten studies which allowed a correction for unreliability in each rating format, only one study (Cotton & Stoltz, 1960) resulted in a correlation between pairs of scores on the same rates below .70 (namely, .63). Four of the ten correlations exceeded 1.00 after correction for attenuation (Bendig, 1954; Campbell, Dunnette, Arvey & Hellervik, 1973; Shapira & Shiuom, 1980; Bernardin, Elliott & Carlyle, 1980). All of these are field studies. Pulakos (1984b) reported an attenuated
correlation of .94 between BARS and BOS in a laboratory study. Only the laboratory studies could provide an accuracy measure and, again, there were generally no differences between formats.

Before we derive the same conclusion regarding formats that is found in almost every review on the subject, let's return to our discussion of possible boundary variables and their effects on research outcomes. We proposed that rating purpose may be an important boundary variable with regard to certain rating formats. In fact, of the forty-two studies that we reviewed involving format comparisons, all but three (Bernardin & Orban, 1984; Sharon & Bartlett, 1969; Taylor & Wherry, 1951) collected performance ratings for research purposes only. As stated earlier, forced choice scales and mixed standard scales were specifically developed to control or eliminate deliberate rating distortion. We located five field studies involving a comparison of the mixed standard format with one or more other formats (Arvey & Hoyle, 1974; Saal & Landy, 1977; Bernardin & Orban, 1984; Dickinson & Zellinger, 1980; Edwards, 1982). Only one of these studies collected ratings in the context of an important purpose (that is, the ratings were used as a source of information for promotions). This single study (Bernardin & Orban) found support for the mixed standard approach.

Our review of the empirical comparisons involving forced choice scales revealed the same methodological consistency across seven field studies. Although it is well known that forced choice scales were developed to control deliberate rating distortion, data collection for all but three of the seven comparisons involving this format was made for research purposes only. The three studies which manipulated the purpose for the appraisal found support for the forced choice method (Sharon & Bartlett, 1969; Sisson, 1948; Taylor & Wherry, 1951). Researchers seem to have lost sight of the fact that forced choice scales are specifically designed to reduce leniency. King, Hunter, and Schmidt (1980), for example, only discuss the method in the context of reducing halo, a variable more closely linked to rater ability rather than to rater motivation (Bernardin & Cardy, 1982).

In addition to generally ignoring the purpose for which appraisal data will or has served, format comparisons face another problem in both the lab and the field. This is the problem of format labeling. Across both lab and field studies, researchers have used more than one format even when claiming to be using only one. For example, Bernardin and Smith (1981) discussed this issue with regard to behaviorally anchored rating scales. The argument applies to other formats as well. A format (for example, BARS, mixed standard) used in one study is often substantially different from the "same" format used in another study. Thus, lab versus field comparisons are confounded by format differences across different research strategies. This confounding applies to all dependent variables under study (Bernardin & Beatty, 1984). For this reason, it is useless to discuss effect sizes, number of significant findings, et cetera as functions of the two research strategies.

Our review of the lab and field research on rating formats reaches a different conclusion than that which is commonly accepted today. Indeed, rating format,
specifically mixed standard scales and forced choice scales, may inhibit or prevent rating inflation which is common in situations where ratings have important administrative significance. Unfortunately, most field research and all laboratory research on rating formats has yet to include this significant boundary variable which could augment or change effects commonly observed in the literature.

Rater Training

Despite the fact that performance appraisal systems of one form or another can be found in 90 percent of American organizations, of the thirty-four rater training studies we located, twenty-six involved student raters in an experimental context (see table 3–1). The majority of both laboratory and field studies found significant effects for rater training. Only three studies were found which involved ratings relevant to actual administrative decisions (Bernardin, 1984; Ivancevich, 1979; Warmke & Billings, 1979). Nineteen of the thirty-four studies, or 56 percent, are laboratory studies involving ratings of hypothetical ratees. From the lab research, we know that we can provide rater training which replaces one response set with another (spread out your ratings across dimensions, for example; Bernardin & Pence, 1980). With the advent of the so-called true scores through the Borman (1978) and Murphy, Kellam, Balzer, and Armstrong videotapes (1982) and the myriad of “paper people,” our new dependent variable is accuracy and we are on a research path which will predictably end up with the “training” comprised of the subjects memorizing each and every performance specimen or, at the least, doing precisely what Borman (1978) had his experts do when deriving “true scores.” It would not be surprising that subjects so trained would rate more like the “experts” than those who were not so trained. The potential generalizability of such training to typical appraisal situations is negligible.

Campbell, Daft, and Hulin (1982) have recommended that we as researchers should avoid asking questions to which we already know the answers and should avoid research tailored for particular research methodologies. It seems we are doing both of these things in our laboratory research methodologies. It is a sad commentary on our discipline that an obvious applied area like rater training should

<table>
<thead>
<tr>
<th>Table 3–1</th>
<th>Training Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Laboratory</td>
</tr>
<tr>
<td></td>
<td>Students/Professors</td>
</tr>
<tr>
<td>Significant effects</td>
<td>11, 16, 18, 31, 42, 48, 51, 52, 57, 58, 64, 69, 73, 74</td>
</tr>
<tr>
<td>Nonsignificant effects</td>
<td>21, 23, 62, 79, 82</td>
</tr>
</tbody>
</table>
be studied almost exclusively with student raters in an experimental context. It is equally sad that our concentration on dependent variables has focused almost exclusively on psychometric characteristics such as halo, leniency, and accuracy, disregarding other more useful applied variables (for example, rater agreement with the rating, perceived fairness in the process, rater evaluations of the training).

Also worth noting is the absence of concern over motivational variables. Virtually every rater training study has concentrated on the rater's ability to rate accurately (for example, if we make raters aware of halo error, they commit less; if we have them practice and become familiar with a difficult rating form, they ultimately rate more accurately). It takes no great insight to realize that, even if the content of the training program is superb, it will not necessarily result in better raters if it fails to motivate them to either acquire the necessary skills or use them. That ability and effort combine to produce behavior has been recognized for years in the performance literature (Campbell, Dunnette, Lawler, & Weick, 1970). We need to pay closer attention to the implications of this relationship for rating accuracy. (See Bernardin & Beatty, 1984, pp. 267–276 for a discussion of possible research avenues.)

As we noted earlier in the chapter, purpose could act as a boundary variable in an appraisal situation. With regard to training, purpose may be a significant moderator of relationships initially derived in controlled laboratory settings. Of the thirty-four studies that dealt with the effects of training, only three of these involved real data, severely limiting any conclusions at this time. Only one study (Bernardin, 1984) found effects of the training in an administrative context. This implies that the motivational effect of rating purpose may be of greater magnitude than the ability-directed and motivationally deficient characteristics of virtually all training interventions currently under study. Such training appears to have great applicability for assessment centers, work samples, and interviewing where presumably rater motivation to rate accurately would be high and relatively stable across raters. However, if the interest of researchers in rater training is with generalizability to appraisals of actual organizational behavior, it is time to reassess the laboratory paradigms which are currently dominating this area of research. Such approaches should also incorporate other potential boundary variables such as realistic time manipulations and rater/ratee personal contact.

Appraisal Purpose

Our review of the empirical literature on purpose of appraisal identified some features of this work that make it an outstanding example of contemporary research on performance appraisal. For example, we found that research on this topic has used the more sophisticated data collection and analytic strategies at our disposal (for example, Williams, DeNisi, Blencoe & Cafferty, 1985; Zedeck & Cascio, 1982). Also, the fact that recent investigations have been conducted is laudable in light of the apparent neglect of this topic just a few years ago. On the other hand,
we also found that manipulations of the purpose of appraisal have been simple one-liners providing abbreviated information to the subject as to the purpose of the appraisal (for example, Bernardin, 1978; Murphy et al., 1982; McIntyre, Smith, & Hassett, 1984; Zedeck & Cascio, 1982). We found no attempts to more realistically engage the subject through the purpose manipulations. Table 3-2 presents a summary of the sixteen studies we identified involving the purpose of appraisal data. Despite the simplistic manipulations, the majority of both lab and field studies found significant effects for the purpose of appraisal data. The laboratory studies indicate that raters weight performance data differently as a function of appraisal purpose (for example, Bazerman & Atkin, 1982; Williams et al., in press). The field studies indicate consistently greater leniency for appraisal data when the data are used for more important personnel decisions such as promotions and retention (for example, Bernardin & Orban, 1984; Taylor & Wherry, 1951). The laboratory studies using leniency as the dependent variable have not supported this finding, however (Bernardin, Abbott & Cooper, 1984; McIntyre et al., 1982; Murphy et al., 1982). This is not at all surprising given the manipulations made in the laboratory studies. We support the views of Berkowitz and Donnerstein who argued that "the meaning the subjects assign to the situation they are in . . . plays a greater role in determining the generalizability of an experimenter's outcome than does the sample's demographic representatives or the setting's realism" (1982, p. 249). With regard to purpose manipulations, we believe the meaning student subjects assign to such a situation does not even approach that which is attached to actual appraisal situations where, for example, a coworker's, subordinate's, and/or acquaintance's promotion, merit increase, or retention may be on the line. Laboratory research should now be directed at interactive variables in the context of more psychologically realistic purpose manipulations. For example, individual difference variables have great potential in this line of research (Bernardin & Beatty, 1984). Cummings (1983) has argued that if trust and commitment characterize the sentiments of employees toward the organization, performance evaluation systems will place greater emphasis on employee performance and career development than on employee control. As noted by Cummings, the extent to which performance evaluation is used as a control device in organizations

<table>
<thead>
<tr>
<th>Table 3-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studies of the Purpose of Appraisal</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Significant effects</td>
</tr>
<tr>
<td>Nonsignificant effects</td>
</tr>
</tbody>
</table>

*Nonsignificant effects with leniency or accuracy as the dependent variable.
depends on the level of trust between the employed and employer as well as the commitment of the employee to organizational effectiveness and the commitment of the employer to employee career development.

This aspect of trust has also been implicated in the work of Bernardin and his associates in which they have administered "Trust in the Appraisal Process Survey" to raters (for example, Bernardin & Orban, 1984). This research found preliminary support for the role of trust in rating distortion. Raters who report less trust in the accuracy of the appraisal process also rate more leniently. This behavior on the part of raters may be an attempt to exercise some control over what they perceive to be an arbitrary and capricious system of evaluation. More creative laboratory paradigms would certainly allow for investigations of these potentially interactive variables.

A Proposed Modal Criterion Setting

Cook and Campbell (1976) and Tunnel (1977) have discussed techniques for increasing external validity. One model recommended has been termed "generalization to modal instances" (Cook & Campbell, 1976, p. 286). This practical model requires a specification of respondents, work settings, and conditions of testing which are modal for a given research area. Let’s say we’re trying to develop a model of the performance appraisal process. External validity for our model would be quite low if student ratings of either professors or hypothetical vignette ratees were used as the method of study. Other than generalizing to somewhat different situations involving student ratings of professors, one would have to seriously question the generalizability of such research. However, a model or theory of the appraisal process which has been tested under conditions most common for appraisal should be more generalizable. Emphasis in this approach should, of course, be placed on common psychological variables with the importance of "mundane realism" dependent on the psychological compatibility of the testing and criterion settings, that is, to the extent that mundane realism augments psychological realism, it is important. A number of practitioners were surveyed for purposes of developing a modal appraisal situation. If one then chooses to adopt this practical model for enhancing external validity, the modal appraisal characteristics will be available.

A questionnaire was prepared in order to derive a modal criterion setting. Questions were written for each of the fifteen relevant parameters of a performance appraisal system identified by Bernardin and Beatty (1984). In addition, practitioners were asked to indicate for which purposes the appraisal data serve (Bernardin & Beatty, 1984, p. 8). A total of 125 personnel administrators who were members of either American Society of Personnel Administrators or the International Personnel Management Association completed the survey. The public and private sectors were equally represented. Respondents were asked to indicate
the most common characteristics of performance appraisal systems within their organization. The survey was mailed to a random sample of 150 members of each association. Time precluded a larger scale study. Given the small sample used, the proposed modal criterion setting presented should be interpreted with caution. The results of the survey are summarized in Table 3-3. Despite all the hoopla over behavioral rating scales, trait ratings are apparently still the most common format across both public and private sector organizations. Ratings are used for important personnel decisions, although public sector respondents indicated that step increases were for the most part automatic regardless of the type of rating form used.

This modal criterion setting is proposed as a frame of reference for assessing the generalizability and potential generalizability of appraisal research. The model is not by any means a recommended system of appraisal. In fact, there are several modal characteristics which are anathema to current academic recommendations. However, we can predict that the greater the compatibility of the research setting with the modal criterion setting, the greater the probability that a finding in the research setting will generalize to a greater number of real appraisal situations. Certainly some of the modal characteristics can be considered more essential than others. Thus, in the construction of laboratory paradigms or the selection of field settings, researchers concerned about generalizability should attempt to hold constant or manipulate such characteristics in their research. For example, a laboratory finding that BARS ratings were less lenient than trait ratings would have very

Table 3-3
Proposed Modal Criterion Setting

1. A supervisor evaluates the performance of six subordinates on an individual basis.
2. Ratees sign off on the rating form and indicate whether they agree or disagree with the rating.
3. A supervisor's ratings are signed off at the next highest level. A supervisor's boss rarely disagrees.
4. Ratees are aware of how fellow workers are evaluated.
5. Ratings are done on an annual basis.
6. Ratings are used as a basis for important personnel decisions (such as step increases and merit pay).
7. Raters feel they don't have adequate time in order to do good performance appraisals.
8. Supervisors are responsible for subordinate work products, that is, supervisory evaluations are affected by subordinate performance.
9. Supervisors have performed the major tasks of the subordinates' jobs.
10. Important countable results are unavailable.
11. Raters feel they themselves rate more accurately than other raters.
12. Raters feel uncomfortable doing appraisals.
13. Supervisors do not feel that formal performance appraisals play a major role in performance feedback for subordinates.
14. Raters evaluate ratee traits (such as dependability, attitude, and initiative).
15. Ratings are formally made relative to scale content, not persons.
little potential for generalizability when modal characteristics 2, 4, 6, 7, 8, and 11 are not a part of the laboratory conditions. To the extent that these characteristics are either manipulated or held constant in the laboratory setting, there is a greater probability that the finding will generalize. Of course, the key here is the psychological fidelity of the research setting and the criterion setting rather than the literal similarity. It is thus not essential that rates sign off on a rating form (modal characteristic #2). It is probably essential, however, that raters believe that ratees will see the resultant ratings. Results from the survey to be discussed in the next section should help identify the essential psychological variables of interest.

Factors Affecting Appraisal Accuracy

Our survey work also revealed the extent to which practitioners believe rating accuracy is affected by various appraisal factors. We content analyzed narratives describing the various factors affecting rating accuracy and found twenty relatively independent factors. We then asked another group of supervisors, subordinates, and administrators to indicate the extent to which they felt each of the factors was affecting rating accuracy. We then compared this result to our content analysis of the research in terms of these factors, that is, whether a factor was manipulated, held constant, or even considered.

The narratives were written by forty-two supervisors representing twenty-seven private and eight public organizations, seventy personnel administrators representing thirty-one private and twenty-four public organizations, and ninety-six subordinates representing twelve private and eight public organizations. Unfortunately, this sample can in no way be considered random. It represents participants from several seminars, responses to a questionnaire mailed to members of the American Society of Personnel Administrators, and future raters in projects related to performance appraisal. Supervisors and personnel administrators also wrote narratives regarding their experiences as ratees. A total of 155 useful narratives were written in response to the following question:

Based on your own personal experience, what factors contribute the most to inaccuracy (accuracy) in performance appraisal? Write a description of a first hand experience you have had as a performance rater, an administrator of an appraisal system, or as the object of a performance appraisal which illustrates what you view as the major reasons for the inaccuracy (accuracy) in appraisal.

The 155 narratives were then clustered into categories by two graduate students and a separate Likert-type item was written for each cluster (see table 3–4). A survey was then conducted to measure the extent to which administrators, raters, and ratees felt each of the factors contributed to appraisal inaccuracy in their organization. Table 3–4 also presents the mean ratings for each of the factors. The surveys were completed by a nonrandom sample of supervisors, personnel administrators, and subordinates who had participated in one or more seminars in personnel management or who were students in an MBA program. A total of fifty,
### Major Factors Affecting Appraisal Inaccuracy

What extent is each statement related to appraisal inaccuracy in your organization?

1. Not at all
2. To a limited extent
3. To some extent
4. To a significant extent
5. To a great extent

<table>
<thead>
<tr>
<th>Statement</th>
<th>Supervisors N = 85</th>
<th>Administrators N = 38</th>
<th>Subordinates N = 194</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters have personal prejudices against certain employees.</td>
<td>2.5</td>
<td>2.7</td>
<td>3.5</td>
</tr>
<tr>
<td>Raters don't have enough time to devote to performance appraisal.</td>
<td>3.8</td>
<td>2.9</td>
<td>2.5</td>
</tr>
<tr>
<td>Raters do not consider performance appraisal to be important relative to other duties.</td>
<td>3.2</td>
<td>3.9</td>
<td>2.7</td>
</tr>
<tr>
<td>Raters are unaware of or do not consider performance factors beyond the control of the employee.</td>
<td>2.4</td>
<td>2.8</td>
<td>4.0</td>
</tr>
<tr>
<td>Raters do not sample a sufficient amount of ratee behavior.</td>
<td>2.2</td>
<td>3.0</td>
<td>3.2</td>
</tr>
<tr>
<td>Raters do not understand the employee's job.</td>
<td>2.3</td>
<td>2.5</td>
<td>3.6</td>
</tr>
<tr>
<td>Raters rate higher than deserved to please certain employees.</td>
<td>3.4</td>
<td>3.6</td>
<td>4.4</td>
</tr>
<tr>
<td>Raters rate higher than deserved to promote a troublesome employee.</td>
<td>2.4</td>
<td>2.1</td>
<td>1.5</td>
</tr>
<tr>
<td>Raters rate higher than deserved because they prefer to avoid confrontations.</td>
<td>3.9</td>
<td>3.8</td>
<td>2.2</td>
</tr>
<tr>
<td>Raters rate higher than deserved because there is no one telling them to rate lower.</td>
<td>2.5</td>
<td>3.9</td>
<td>2.7</td>
</tr>
<tr>
<td>Raters rate higher than deserved because of friendship.</td>
<td>3.0</td>
<td>3.3</td>
<td>3.4</td>
</tr>
<tr>
<td>Raters rate lower than deserved in order to get rid of a troublesome employee.</td>
<td>2.1</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Raters feel uncomfortable evaluating other people.</td>
<td>3.7</td>
<td>3.6</td>
<td>2.4</td>
</tr>
<tr>
<td>Raters have ambiguous criteria on which to rate (for example, personality traits).</td>
<td>3.2</td>
<td>4.2</td>
<td>3.6</td>
</tr>
<tr>
<td>Raters do not consider all important aspects of an employee's job.</td>
<td>2.4</td>
<td>2.9</td>
<td>3.2</td>
</tr>
<tr>
<td>Raters are not rewarded for accurate appraisals.</td>
<td>3.4</td>
<td>4.0</td>
<td>3.3</td>
</tr>
<tr>
<td>Raters avoid the whole responsibility if they possibly can.</td>
<td>2.6</td>
<td>2.9</td>
<td>2.2</td>
</tr>
<tr>
<td>Raters forget how employees performed early in the appraisal period.</td>
<td>2.3</td>
<td>3.1</td>
<td>2.8</td>
</tr>
<tr>
<td>Raters weight how the employee performed late in the appraisal period more heavily than early performance.</td>
<td>2.4</td>
<td>3.1</td>
<td>2.9</td>
</tr>
<tr>
<td>Raters attach too much weight to relatively unimportant performance factors.</td>
<td>2.5</td>
<td>3.0</td>
<td>3.7</td>
</tr>
</tbody>
</table>
twenty-nine, and sixty-eight different private and public sector organizations were represented in responses by the supervisors, administrators, and subordinates, respectively.

We have drawn several implications from these results with regard to essential research needs in performance appraisal. As we see them, the major implications are:

1. Supervisors and administrators feel factors related to a deliberate motivation to distort ratings have more to do with inaccuracy than the ability to rate. Subordinates felt that factors related to the rater's knowledge about the rated job were relatively more important. The importance attached to rater motivational factors by raters and administrators suggests a need to study this general source of rating variance and to investigate methods to control deliberate distortion, including rater training. Our review found very few studies which addressed the issue of deliberate rating distortion. One recent study which investigated deliberate rating distortion by raters found strong effects (Bernardin & Orban, 1984). The survey results also suggest we should look more carefully at rating formats designed to control rating distortion, such as forced choice.

One boundary variable identified as critical in terms of leniency is the extent of positive consequences for the rater who inflates ratings. An inflated rating can be predicted, for example, if the consequence of such a rating is to promote an obnoxious subordinate into a new division. Under those conditions, a training program aimed at enhancing observational skills will probably not inhibit leniency error or increase accuracy. Kane and Lawler (1979) discussed the issue of rater motivation in the context of the "tendency to dissemble." They viewed this tendency as a function of the probability of being detected in making invalid ratings, the expected value to the rater of making invalid ratings, and the expected disincentive value to the rater if invalid ratings are exposed. Kane (1980) has developed these relationships into their expectancy components. This suggests a study in which leniency is regressed on the expectancy components. We were unable to identify any studies which investigated this relationship. Our survey results indicate practitioners would be most interested in such a study and, of course, realistic interventions which could reduce the tendency to dissemble.

2. The greatest difference in responses between supervisors and subordinates concerned the extent to which raters adequately consider factors affecting employees which are beyond the employees' control (item 4 in table 3-4). Subordinates thought this was a source of inaccuracy to a "significant extent" while supervisors thought it was a problem only to a "limited extent." This finding is of course in line with applied predictions in attribution theory (see chapter 4 of this book) and suggests a rater training program which focuses on identifying potential constraints on individual performance and defining and
agreeing with the subordinate on the boundaries of causality as much as possible prior to the appraisal period. Subordinate agreement with the supervisor’s appraisal would be a useful dependent variable. It would not be difficult to conceptualize a laboratory setting which could be used to unambiguously test such an approach to training. The boundary variable that should be manipulated in the laboratory is the ambiguity in attributions regarding individual performance and the alleged constraints. In a field study, Bernardin (1984) successfully reduced disagreements in self versus supervisory appraisals as a function of this training.

3. Subordinates also felt that supervisors did not really understand the subordinates’ jobs (see items 5, 6, 14, 15, and 20 in table 3–4). For example, supervisors do not weight the various job factors appropriately in making their ratings. More specifically, a rater may attribute more importance to the successful completion of some aspects of the job while discounting the importance of others. That is, the job and its various components may be redefined in a personal manner by the incumbent. Some aspects of the job are viewed as more important, perhaps due to the ability of the incumbent to perform them, while other aspects of the job are discounted because the incumbent does not perform them as well. It is likely that such redefinitions and attributions of job criticality may lead to different rater strategies with some raters giving more weight to the successful performance of some aspects of the job, and other raters viewing other tasks as more crucial. These redefinitions and attributions of criticality, as noted above, may stem from the rater’s own ability or previous success at performing the tasks required.

No research has yet been conducted to investigate this phenomenon. Attempts have been made, however, that suggest this may be a worthwhile area of study. For example, Mitchell and Kalb (1982) showed that supervisors who have had experience with a subordinate’s job tend to attribute poor performance to external factors. However, supervisors with little experience on the job may attribute the performance of a poorly performing subordinate to internal factors in accordance with the actor-observer bias commonly found in attribution studies. For supervisors who make more attributions to external causes for poor performance, the performance appraisals may be more lenient; likewise, supervisors with no experience may be more severe raters. Also, in the former condition, there would probably be greater agreement between rater and ratee on the rater’s appraisal.

Little research has been directed at discovering the role of attributions for the performance of others on performance appraisal. What is commonly assumed is that a ratee is evaluated on a continuum of success or proficiency on the job. Process variables such as attributions for this level of performance are often neglected, leading to confusion as to whether the rater is rating performance per se, or other attributes of the ratee such as effort or ability.

Now that we have recognized this omission in performance appraisal research, how might we study it? One possible way is to use a paradigm similar to that of
Mitchell and Kalb (1982), in which raters are first provided with experience on the job on which they will be rating subordinates' performance. Some supervisors may be "held out," that is, not given any experience with the job. There will have to be some inducement in the form of a monetary incentive to realistically involve the raters. Further, unlike the Mitchell and Kalb paradigm, several levels of performance outcomes may be desired, as well as several levels of supervisory experience with the subordinates' job.

More important, and certainly more difficult to achieve in the lab, is the definition of task criticality mentioned above. Such a redefinition of task criticality will require an extended period of incumbency on the part of the rater. This may preclude a direct test of the effects of rater experience in the lab. Perhaps one means would be to explore this in a field setting. A rater's past performance record on the job that he now supervises may be used as an independent variable in the ratings of subordinates. Also, different levels of performance on the various dimensions may prove to be useful covariates that can be used to obtain an unbiased appraisal of subordinate performance, one free of rater idiosyncracies with regard to differential weighting of performance. Since many U.S. companies use internal promotions as a standard practice and many also use task-based appraisal systems, a field test of the above would be very useful providing sufficient variance is available.

Discussion

Our literature review and two surveys lead to several suggestions for future research. As stated earlier, forced choice scales and mixed standard scales were specifically developed to control or eliminate deliberate rating distortion, a condition indicated in our survey to be most evident when appraisal data were used for important decisions (a modal characteristic). Is it possible to study the effects of these formats in the laboratory with students in such a way that the results may be generalizable to the modal criterion setting? With regard to the purpose boundary variable, the typical method for manipulating purpose of the appraisal is unacceptable. In general, this method has amounted to a single line in the directions provided to raters (for example, these ratings will be used as feedback to the instructor; they will be used to decide whether this instructor receives a merit increase). Our survey data indicated that raters typically rate with the knowledge that they will interact with the ratee subsequent to the rating process. All of these factors should affect the tendency to dissemble and thus should be either manipulated or held constant in a format comparison involving a method specifically designed to control that tendency.

It is possible to control these aspects of performance which are to be rated (for example, by videotape) while still providing an experimental context involving some or all of the modal characteristics discussed above. Thus, one or more instructors’
lectures could be videotaped and the same person could interact before and after
the viewing of the videotape in such a way that all of the critical modal variables
could either be manipulated or held constant. Ratings from forced choice scales
should then be more accurate (using "true scores") and less lenient than more
transparent scales. Such a design would provide either a between- or within-
subjects comparison of the two supposedly nontransparent formats as well.

What external validity might this have? We believe a great deal. Florida and
Tennessee have passed legislation mandating state-administered performance-
based merit pay systems for their teachers. Given cost considerations, the only
realistic rating source for such appraisals is probably in the individual schools (for
example, principals, assistant principals), where the tendency to dissemble would
be great. A rating format which can control this inevitable problem would undoubt-
dedly be welcomed by state administrators concerned about rating inflation,
where numbers are to be linked to monetary allocations.

Regarding rater training, our literature review indicated that training efforts
typically focused on providing raters with an opportunity to acquire better rating
skills while disregarding the raters' motivation to acquire and/or use the skills.
Again, a motivational component that could act as a critical boundary variable
seems to have been omitted in our research. What form might research take in in-
vestigating this critical variable? One suggestion we make is to vary the amount of
disincentive to dissemble. This could be done by using a policy-capturing ap-
proach to understand raters' weighting of performance information and conse-
quent biases. For example, raters could participate in an organizationally spon-
sored program designed to uncover rater inconsistency and inaccuracy. Via video-
tape raters would be presented with orthogonally varied performance outcomes on
various dimensions. The profile of each rater's performance-rating policy could be
reviewed in a separate meeting between the rater and his immediate supervisor.
Implementing such a program allows the identification of rater idiosyncrasies and
the communication to the rater that the organization has the capability to discover
rating bias. Bernardin and Beatty (1984) have suggested other potentially useful
areas which are amenable to both laboratory and field methods.

Summary

We find that the fit between the research literature and the modal criterion setting
of our survey can be likened to that of fitting square pegs into round holes. There
appears to be an increasing emphasis on methodologically unconfounded, inter-
nally valid laboratory and field research, the results of which have added little to
our understanding of performance appraisal. In the immortal words of Robert
Wherry, "We don't know what we are doing, but we are doing it very carefully
and hope you are pleased with our unintelligent diligence" (1957, p. 1). We
should add that we are also doing it very quickly since such research is so easy to
conduct. The research paradigms currently in vogue establish an appraisal situation which is far more like assessment in assessment centers, or work samples, or ratings of employment interviewees. There is also considerable overlap between this type of performance appraisal research and laboratory research on leadership (Bernardin & Beatty, 1984). Yet there is little attempt to merge the research literature.

A modal criterion setting has been proposed as a framework for future research methodologies on the general type of performance appraisal conducted in over 90 percent of U.S. organizations. Certainly a more systematic and detailed delineation of the modal setting is in order. Another survey has identified several critical issues related to appraisal which have been largely ignored in both laboratory and field studies. The results of this and similar surveys should direct applied researchers to areas most in need of "application" and help in the development of more psychologically realistic research settings.

References


62 • Generalizing from Laboratory to Field Settings


