Psychological Bulletin
1980, Vol. 87, No. 1, 72–107

Performance Rating

Frank J. Landy and James L. Farr
Pennsylvania State University

The most ubiquitous method of performance appraisal is rating. Unfortunately, ratings have been shown to be prone to various types of systematic and random error. Studies relating to performance rating are reviewed under the following headings: roles, context, vehicle, process, and results. In general, cognitive characteristics of raters seem to hold the most promise for increased understanding of the rating process. A process model of performance rating is derived from the literature reviewed. Research in the areas of implicit personality theory and variance partitioning is combined with the process model to suggest a unified approach to understanding performance judgments in applied settings.

The measurement of performance in industrial settings has occupied the attention of psychologists for 50 years. Performance description and prediction plays an important role in all personnel decisions. Criteria are necessary for validation studies and training evaluation; indices of effectiveness or relative worth are necessary for administrative decision making with respect to current employees; performance-related information is necessary for feedback and employee counseling; there is even some indication that the process of performance evaluation may function as a reward and be capable of inducing feelings of satisfaction in some employees (Landy, Barnes, & Murphy, 1978).

Unfortunately, realizing the importance of performance measurement and actually measuring performance accurately are two different matters. In some ideal sense, complete performance measurement would include the combination of objective, personnel, and judgmental indices (Landy & Trumbo, in press). Unfortunately, it is difficult to obtain objective indices of performance for many job titles. In addition, personnel information is applicable to a small portion of the employee population in any organization (e.g., 5% of the employees may have 100% of the accidents, less than 8% of the employees may have more than one unexcused absence per year, tardiness records are not well kept, etc.).

Consequently, most individuals concerned with performance measurement depend on judgmental indices of one type or another. Guion (1965) reported that 81% of the published studies in the *Journal of Applied Psychology* and *Personnel Psychology* between 1950 and 1955 used ratings as criteria. Blum and Naylor (1968) sampled articles from the *Journal of Applied Psychology* for the period from 1960 to 1965 and found that of those using criterion measurement, 46% measured performance via judgmental indices. Landy and Farr (1976) reported that 89% of 196 police departments in major metropolitan areas used supervisory ratings as the primary form of performance measurement. Finally, Landy and Trumbo (in press) reported that a literature review of validation studies in the *Journal of Applied Psychology* between 1965 and 1975 revealed that ratings were used as the primary criterion in 72% of the cases. By any standard, judgmental measurements of performance are widely used.

In spite of the widespread use of judgmental indices of performance, there has been a constant dissatisfaction with these measures on the part of both researcher and practitioner. The source of this dissatisfaction has been

We are indebted to Janet Barnes for help with the literature review.

Requests for reprints should be sent to Frank J. Landy, Department of Psychology, Pennsylvania State University, University Park, Pennsylvania 16802.

 0033-2909/80/8701-0072$00.75

the vulnerability of these measures to both intentional and inadvertent bias. As a consequence, an enormous amount of research has been conducted in an attempt to improve the validity of judgmental indices of performance. These studies have covered a wide variety of issues, such as rater and ratee individual differences, types of formats, conditions surrounding the judgmental process, and so forth. In this article, we review the outcomes of this research.

We limit the scope of this review to a consideration of one particular form of performance judgment—the performance rating. We choose to concentrate on this method for three reasons: (a) As indicated in earlier reviews (Guion, 1965; Landy & Farr, 1976; Landy & Trumbo, in press), the rating is by far the most ubiquitous form of performance judgment; (b) research on various aspects of rating is more common than research on any other judgmental index of performance, and (c) other judgmental methods, such as ranking, pair comparison estimation, and other forms of worker-to-worker comparison, imply a qualitatively different discrimination process. Thus, the review deals primarily with a consideration of rating methods. In addition, since Wherry (Note 1) completed an exhaustive review of performance rating research prior to 1950, we deal predominantly with the literature appearing subsequent to that review. Other reviews have appeared since 1950 (e.g., Barrett, 1966b; J. P. Campbell, Dunnette, Lawler, & Weick, 1970; Lopez, 1968; Miner, 1972; Smith, 1976), but these have generally been more narrowly focused than is the present review.

We exclude detailed consideration of several broader issues in the measurement of job performance. These include such issues as the dynamic nature of criteria (B. M. Bass, 1962; Ghiselli & Haire, 1960; Prien, 1966), the composite criterion versus multiple criteria controversy (Dunnette, 1963; Guion, 1961; Schmidt & Kaplan, 1971), and the relationship of ratings to more general theories or models of human performance (James, 1973). These are excluded because they deal with all forms of criterion measures, not just ratings of performance. Their inclusion would necessitate extensive space that is not available, and their discussion would not be appropriate here. The relationships between performance rating and broader organizational questions of selection, training, counseling, and job satisfaction are generally not addressed. Works by Barrett (1966b), Lopez (1968), and Miner (1972), among others, address many of these relationships.

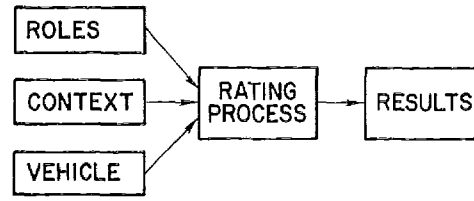

Figure 1. Component model of performance rating.

The literature in the area of performance rating is fragmented. Some do research on different rating formats, whereas others examine characteristics of raters and ratees. This insularity has tended to obscure the fact that performance rating is best thought of as a system comprising many different classes of variables. The rating instrument and the characteristics of raters and ratees are only parts of that larger system. At a general level, that system might include the following classes of variables: (a) the roles (rater and ratee), (b) the vehicle (the rating instrument), (c) the rating context (the type of organization, the purpose for rating, etc.), (d) the rating process (administrative constraints, individual rater strategies, etc.), and (e) the results of the rating (raw and transformed performance information, actions based on that information, etc.). Our review deals with research results bearing on those major classes of variables. Figure 1 is a graphic representation of how those components might interact. Although Figure 1 might be heuristically helpful in structuring the body of the review, it is not particularly illuminating with respect to the specific influences affecting the judgment that one person makes of another's performance. Consequently, after reviewing the available research evidence, we present a more elaborate model of the performance rating process that should be more theoretically useful than Figure 1.

There are some unique constraints on the performance rating literature. Consider the

phenomenon in question. Typically, a supervisor is asked to consider the past performance of one or more subordinates. The period of past performance is typically 1 year. During that period of time, the supervisor and the subordinate have interacted frequently and probably know each other reasonably well. Most research on the topic of performance rating has considered the adequacy of rating in this type of context. For that reason, we have excluded from consideration studies dealing with the evaluation of applicants in interview situations. This literature has been well covered elsewhere and is not directly relevant to traditional performance rating. We conceive of the prototypic performance rating as a retrospective synthesis by one individual of the efforts or performance of another. Thus, we are dealing with an appraisal of a long string of actions rather than a single one; in addition, we are dealing with a constellation of activities rather than with single physical or mental operations in isolation.[1] We have included in the review studies of simulated work settings that have experimentally manipulated variables of interest to performance rating, if the behaviors of the ratees were observable rather than represented only by a pencil-and-paper task.

Since the research on performance appraisal is predominantly in the form of field studies, we are unable to make comparisons from one study to another with respect to the dimensions of performance examined. Each organization has a different idea of what may be important in assessing their people; consequently, each rating instrument is ultimately unique. This is unfortunate, since there is every reason to believe that different types of performance may be evaluated more or less accurately than others. For example, interpersonal skills might be more accurately evaluated than creativity. However, this is a source of variation over which we have no control as reviewers. These are the structural boundaries of the review.

Roles

Rater Characteristics

A great deal of research has examined the relationship between characteristics of the rater and various criteria of rating effectiveness. These studies are grouped into three classes: personal characteristics of the rater, type of rater vis-à-vis the ratee, and rater knowledge of the ratee and the job.

Personal characteristics. Among the personal characteristics of the rater that have been investigated as possible sources of rating variation are demographic variables, psychological variables, and job-related variables.

The most popular demographic variable has been the sex of the rater. These studies have all been published since 1970. In the majority of these studies, there has been no consistent effect of rater sex on ratings obtained in various contexts, including instructional settings (Elmore & LaPointe, 1974, 1975; Centra & Linn, Note 2), simulated work settings (Rosen & Jerdee, 1973), and laboratory research settings (Jacobson & Effertz, 1974; Mischel, 1974). Lee and Alvares (1977) found that rater sex affected the description of supervisory behavior but not the evaluation of such behavior. In a simulated work setting, London and Poplawski (1976), studying college students' evaluations in simulated appraisal and interview situations, found that female subjects gave higher ratings on some dimensions but not on overall performance. Hamner, Kim, Baird, and Bigoness (1974) found that females gave higher ratings than did males when evaluating performance in a simulated work setting, especially for high levels of performance.

Several studies have examined the effect of the race of the rater on ratings. Research conducted by the Educational Testing Service in conjunction with the U.S. Civil Service Commission (Crooks, Note 3) found that in a majority of cases, supervisory raters gave higher ratings to subordinates of their own race than to subordinates of a different race. Hamner et al. (1974) found results similar to those reported in Crooks, but the effect accounted for only 2% of the rating variance. DeJung and Kaplan (1962) and Cox and Krumboltz (1958) obtained results comparable to those of Crooks, with stronger effects for black raters. Schmidt and Johnson (1973), examining peer ratings in an industrial setting, found no race of rater effect. The work

[1] We are grateful to an anonymous reviewer for helping us to elaborate this position.

setting in which this research took place was highly integrated with subjects who had been exposed to human relations training. The ratings were gathered for research purposes only and required a prescribed distribution. It is likely that some or all of these factors influenced the results of this study.

Mandell (1956) and Klores (1966) have examined the effects of rater age on performance ratings. Mandell found that younger supervisors were less lenient in their ratings of subordinates, whereas Klores found no effect of supervisor age in his study of forced-distribution ratings.

The education level of raters was examined by Cascio and Valenzi (1977). They found a significant effect of rater education on supervisory ratings of the job performance of police officers, but the effect accounted for such a small percentage of total rating variance that Cascio and Valenzi concluded that rater education was of no practical importance in their study.

A large number of psychological variables have been studied as possible influences on performance ratings. Unfortunately, in most instances, there has been only a single study investigating any one variable. Thus, general conclusions are difficult, if not impossible, to make about their effects.

Mandell (1956) found that raters who were low in self-confidence were less lenient in their ratings of subordinates than raters high in self-confidence. Lewis and Taylor (1955) reported that individuals high in anxiety tended to use more extreme response categories than those lower in anxiety. Rothaus, Morton, and Hanson (1965) found that increased psychological distance of the rater tended to result in ratings that were more critical and negative.

In a study of rater policy using regression and cluster analytic methodology, Zedeck and Kafry (1977) investigated whether several psychological variables would be related to how a rater used performance information to form overall ratings of performance. Their results indicated that interest measures, social insight, and intelligence measures (verbal or nonverbal) were not significantly related to the rating strategies of the subjects.

Schneier (1977) found that the cognitive complexity of raters had an effect on ratings. Cognitively complex raters were less lenient and demonstrated less restrictions of range with behaviorally anchored scales than did cognitively simple raters. The cognitively complex raters also exhibited less halo in their ratings than did the simple raters, with both behavioral scales and a simpler form of rating scale. Cognitively complex raters also preferred the behaviorally anchored scale to the simpler format.

Among the personal characteristics of the rater that can be thought of as job-related variables are the rater's job experience, performance level, and leadership style. The results of studies that examined the rater's length of job experience are mixed. Jurgensen (1950) found that more experienced raters had more reliable ratings, and Mandell (1956) noted that raters with more than 4 years of experience as supervisors tended to be more lenient in their ratings than were raters with less experience. Klores (1966) obtained no significant effect of rater experience. Cascio and Valenzi (1977) found a significant effect of rater experience but noted that it accounted for only a small percentage of total rating variance.

Several studies have found that the performance level of the rater affects the nature of the ratings assigned to others by that rater. D. E. Schneider and Bayroff (1953) and Bayroff, Haggerty, and Rundquist (1954) reported that peers who received high aptitude test scores and were rated positively during training gave ratings of their fellow trainees that were more valid in predicting subsequent job performance. Mandell (1956) found no difference in central tendency between good and poor job performers but did find that those raters who were poor performers tended to disagree more with consensus ratings of subordinates than did the more favorable performers. Kirchner and Reisberg (1962) found that the ratings given to subordinates by supervisors high in job performance were characterized by greater range, less central tendency, and by more emphasis being placed on the independent action of subordinates as the basis for ratings. In a related study Mullins and Force (1962) obtained evidence for a generalized ability to rate others accurately. Peer raters who were more accurate in judging one skill of their co-

workers also were accurate in judging another performance dimension. (Accuracy was assessed by comparing the ratings with scores on pencil-and-paper tests.)

The effects of the rater's leadership style on the ratings have been examined by E. K. Taylor, Parker, Martens, and Ford (1959). They found that production-oriented supervisors gave lower ratings to subordinates. Klores (1966) reported that raters who were high in consideration were more lenient in their ratings of subordinates than were raters who were high in initiation of structure. Those raters high in initiation of structure exhibited more range in their ratings and gave more weight to the planning and organization function when evaluating their subordinates' overall job performance. Zedeck and Kafry (1977) found that leadership style was not correlated with rater strategies, as identified by regression and cluster analytic techniques.

Type of rater. The studies reviewed in this section are concerned with rating differences obtained with raters who differed in the type of relationship they held in regard to the ratee (e.g., supervisor, peer, self, or subordinate). Studies that focused on only one type of rater are generally not reviewed here. Discussions of investigations that deal with the various single types of raters are available elsewhere (e.g., Guion, 1965; Kane & Lawler, 1978; Lewin & Zwany, 1976).

The most frequent rater type comparison has been that of supervisory rating versus peer rating. These studies have generally demonstrated differences between the two rater types. Springer (1953), Rothaus et al. (1965), and Zedeck, Imparato, Krausz, and Oleno (1974) found that supervisors were less lenient in their ratings than were the peers of the ratees. Klieger and Mosel (1953) and Springer both found that there was more interrater agreement with supervisory ratings than with peer ratings, but L. V. Gordon and Medland (1965) reported greater reliability for peer ratings of leadership than for similar supervisory ratings.

Although Booker and Miller (1966) obtained general agreement between peers' and instructors' ratings of Reserve Officers' Training Corps students, Springer (1953) and Borman (1974) reported less supervisor–peer agreement than was found within either type of rater group. Data reported by Borman (1974) and Zedeck et al. (1974) suggest that supervisor–peer rating differences may be expected and do not necessarily suggest that either type of rating is invalid or unreliable. Borman (1974), as well as Landy, Farr, Saal, and Freytag (1976), found that the dimensions of job performance resulting from the development of behavior-anchored scales for use by peers and supervisors differed. Zedeck et al. obtained similar dimensions of performance for the two rater types, when developing behavior-anchored scales but did find that the specific examples of job behaviors used to anchor the dimensions differed between the two rater types. Thus, supervisory and peer ratings may represent two distinct views of a common individual's job performance and may be equally valid, even though they are not highly correlated.

Supervisory ratings have also been compared with the ratees' self-ratings. Parker, Taylor, Barrett, and Martens (1959) and Kirchner (1965) found that self-ratings were more lenient than supervisory ratings, but Heneman (1974) found less leniency with self-ratings than with supervisory ratings. Heneman's data were gathered for research purposes via a mailed questionnaire that was returned to the researcher. These factors may have affected his results. Kirchner and Heneman reported more halo in supervisory ratings than in self-ratings, whereas Parker et al. found no differences in halo. Both Kirchner and Parker et al. reported only moderate agreement between supervisory ratings and self-ratings.

Lawler (1967) and Klimoski and London (1974) examined self-, supervisory, and peer ratings of performance. Lawler found that supervisory and peer ratings exhibited greater convergent and discriminant validity than did self-ratings. Klimoski and London reported that each rater type was distinct, with regard to use of information, and that supervisory and peer rating strategies were more similar than self-ratings. Supervisory ratings demonstrated a strong correlation between effort and performance ratings, whereas peer ratings and self-ratings differentiated between effort and performance.

A few studies have compared peer ratings to other nonsupervisory ratings. Bartlett

(1959) reported that whereas peer ratings on a forced-choice scale of leadership were useful for both evaluative and diagnostic purposes, self-ratings on a similar scale were adequate only for diagnostic purposes. Centra (Note 4) compared peer and student ratings of college instructors. Peer ratings were found to be more lenient and to have lower interrater agreement than the student ratings.

Freeberg (1969), Fiske and Cox (1960), and Rothaus et al. (1965) compared peer ratings with observer ratings in various role-playing and group activities. Fiske and Cox and Rothaus et al. found that peer ratings were more lenient than observer ratings. Freeberg reported that when peers and observers had similar relevant contact with the ratee, peer ratings were more valid predictors of cognitive skills than were the observer ratings. Kraut (1975) also found that peer ratings in a month-long management training program were more predictive of promotion and future performance appraisals than were ratings by the training staff.

Rater knowledge of ratee and job. Although some minimum rater knowledge of the ratee's job performance and of the job in question is certainly necessary before valid ratings can be obtained, the extent of knowledge that is necessary has been a focus of much research. Several studies have found only low to moderate agreement among the ratings made by supervisors at differing organizational levels, relative to the ratee (Berry, Nelson, & McNally, 1966; Borman & Dunnette, 1975; J. P. Campbell, Dunnette, Arvey, & Hellervik, 1973). Whitla and Tirrell (1953) found that first-level supervisors' ratings more accurately predicted job knowledge test scores of subordinates than did the ratings of second- or third-level supervisors. Zedeck and Baker (1972) reported better construct validity for ratings by first-level supervisors than for those by second-level supervisors. Individuals with more knowledge of the requirements of the particular job have been found to be less influenced by serial position (Wagner & Hoover, 1974) and to more validly predict future performance (Amir, Kovarsky, & Sharan, 1970) than individuals with less knowledge of the job requirements.

The amount and type of contact between the rater and ratee has also been of concern to performance appraisal researchers. Although Ferguson (1949) reported that reliability of ratings increased as the amount of rater-reported acquaintance with the ratee increased, more recent research has not generally supported this finding. Klieger and Mosel (1953) found no effect of rater-reported opportunity to observe the rater on rating reliability. Fiske and Cox (1960), L. V. Gordon and Medland (1965), and Klores (1966) found no effect of length of rater acquaintance with ratee. Hollander (1957, 1965) found no differences in peer rating reliability or validity for ratees who had been acquainted for 3, 6, or 12 weeks. Brown (1968) found that peer raters were not influenced by degree of acquaintance but that untrained peer raters' ratings were characterized by increased halo for less well-known ratees. Waters and Waters (1970), Amir et al. (1970), and Suci, Vallance, and Glickman (1956) reported little or no effect on the validity or reliability of ratings, when the rater's friendship with the ratee was considered. Finally, Freeberg (1969) reported that the relevance of the rater–ratee acquaintance was important in terms of the validity of the ratings. Raters who interacted with the ratees in a situation relevant to the dimension being rated were more valid in their evaluations than were raters who interacted with the ratees in a nonrelevant situation. Similarly, Landy and Guion (1970) reported that raters with daily but peripheral contact with ratees had a median interrater reliability of .24, in contrast to a median reliability of .62 for those raters with more relevant contacts with the ratees. Thus, relevancy rather than frequency of contact appears to be the critical factor.

Summary

The research on rater characteristics provides relatively few general conclusions. Since most studies examine only one or a few characteristics, it is likely that unmeasured or unreported variables have had some effect on the results of any single study. This results in a chaotic pattern of findings in many in-

stances. Nevertheless, some generally consistent effects can be described. Sex of the rater does not generally affect ratings, although female raters may be more lenient. Raters usually give higher ratings to same-race ratees, although this may be moderated by the degree of contact that members of each race have with each other. Rater age and education have been studied too infrequently to make general statements about their effects.

Psychological characteristics of raters have not been systematically researched, but it appears (and has been empirically demonstrated to some extent) that cognitive complexity may be an important variable to examine. There is a large body of research in other content areas which suggest that cognitive complexity affects information processing and evaluation.

Rater experience appears to positively affect the quality of performance ratings, but the mechanism or mechanisms responsible (e.g., more training or experience with the rating form, better observation skills, better knowledge of the job requirements, etc.) is not known. The general job performance of the rater is related to rating quality, with better performers also providing higher quality ratings. Production-oriented (as opposed to interaction-oriented) raters seem to be less lenient and to pay more attention to planning activities.

Comparisons of different types of raters suggest that in general, one should expect only low to moderate correlations among raters of different types (e.g., peer, supervisory, self, etc.). It cannot be stated that any one type of rater is more valid than any other, although peer ratings appear to be especially useful for predicting promotions. Peer ratings appear to be more lenient than supervisory ratings. As Borman (1974) and others have suggested, the best conclusion may be that different types of raters have different perspectives on performance that influence their ratings. Lawler (1967) and Blood (1974) have noted that these differences may provide valuable information for the diagnosis of organizational problems.

Raters require knowledge of the individual ratee and of the requirements of the ratee's job to adequately evaluate job performance. The relevance of the rater–ratee interaction is apparently more important than simply the amount of interaction.

Ratee Characteristics

The research on the effects of the ratee characteristics on performance ratings is divided into two broad categories: personal characteristics and job-related variables.

Personal characteristics of the ratee. The two ratee characteristics that have been examined most frequently in recent research have been sex and race. Much of the research concerned with ratee sex supports the hypothesis that the sex stereotype of the occupation (i.e., whether a particular job is typically perceived as masculine or feminine) interacts with the sex of the ratee. Studies in which the occupation would be likely to be perceived as masculine (e.g., managerial positions) have found that females received less favorable evaluations than did males (Schmitt & Hill, 1977). In addition, Terborg and Ilgen (1975) found in an in-basket exercise that whereas female ratees received ratings similar to those of males, females received lower salaries and less challenging job assignments. Rosen and Jerdee (1973) and Bartol and Butterfield (1976) reported in simulation studies that the sex of the supervisor influenced the rater's perceptions of the appropriate behavior of the supervisor in a sex stereotypic fashion. Elmore and LaPointe (1974, 1975) found that students gave essentially equal ratings to male and female college instructors, an occupation perhaps perceived as less sex specific than management jobs. Lee and Alvares (1977) obtained no effect of ratee sex on evaluations of interviewers. Once again, perhaps the job of interviewer is considered to be neither masculine nor feminine. Bigoness (1976) and Hamner et al. (1974) examined ratee sex effects in semiskilled and low-skilled tasks. Bigoness and Hamner et al. both found that females received higher ratings than did males in a simulation study in which objective performance was controlled. Again, since sex stereotypes were not measured, it is difficult to determine if these studies support the interaction hypothesis. Jacobson and Effertz

(1974) obtained results opposite to those predicted by the sex role stereotype hypothesis. They found that male leaders were evaluated more negatively than were female leaders but that male followers received higher ratings than did female followers. It should be noted that many of the studies that examined the effects of ratee sex on evaluations were simulations. Relatively few studies (Elmore & LaPointe, 1974, 1975; Schmitt & Hill, 1977) have been conducted in which real-world performance of the ratee was being rated.

The effect of the race of the ratee has been examined in several studies. Most of these investigations have used ratings of the real-world performance of ratees as the behavior of interest, whereas some have used simulation methodology. Ratees have been found to receive higher ratings from same-race raters by Crooks (Note 3), DeJung and Kaplan (1962), and Hamner et al. (1974), whereas Schmidt and Johnson (1973) found no such effect with peer ratings that were obtained in a highly integrated setting. Landy and Farr (1976) reported that supervisors (the large majority of whom were white) rated the performance of white police officers more favorably than that of black officers on four of eight rating dimensions.

Other studies have demonstrated an interaction between ratee race and ratee performance level. Bigoness (1976) and Hamner et al. (1974), both using videotaped task performance controlled for level, found interactions of race and objective performance levels. Bigoness reported that among low performers, blacks were rated more favorably than were whites, whereas there were no racial differences for the high performers. Hamner et al. found that raters significantly differentiated between high and low white performers but did not for black ratees.

Huck and Bray (1976) and Schmitt and Hill (1977) both examined ratings gathered in assessment center settings. Huck and Bray found that black female assessees received lower ratings than did white female assessees. The validities of those ratings for predicting future job performance were about equal for blacks and whites. The black women also received lower criterion ratings than the white women. Schmitt and Hill reported that black female assessees tended to receive lower ratings when their assessment center group was composed principally of white males than when the group was better integrated in terms of race and sex.

Several studies that were primarily interested in the validity of selection devices for black and white workers have reported data for performance ratings for the racial subgroups. Farr, O'Leary, and Bartlett (1971) found that white employees received higher performance ratings than blacks in 13 of 22 comparisons. The other 9 comparisons revealed no differences in the rating means for the two groups. Greenhaus and Gavin (1972) reported that white employees were rated higher than blacks on all three supervisory ratings used in their study. Toole, Gavin, Murdy, and Sells (1972) split their workers into younger and older subgroups, the cutoff being age 35. There were no racial differences on a rating measure for the older workers, but white workers received higher ratings than did blacks in the younger group. Kirkpatrick, Ewen, Barrett, and Katzell (1968) found only one significant difference among 8 possible comparisons of rating means for black and white workers. In that one case the white workers received a higher rating than black workers. Crooks (Note 3) reported that white employees were rated more favorably than were black employees but that white employees also received a higher mean score on an objective test of job knowledge. Fox and Lefkowitz (1974) found no mean racial difference for a supervisory rating measure.

A. R. Bass and Turner (1973) found no significant mean differences for black and white raters, when age and job tenure were held constant for full-time employees, and they found small but statistically significant racial differences (with white ratees evaluated more favorably) for part-time workers. However, ratings of black and white employees were differentially related to more objective criterion measures. The ratings of black employees were more strongly related to attendance and error data than were those of white employees. Crooks (Note 3) reported that black ratees received more valid ratings from black and white raters. Validity of the ratings was measured by their relationship to scores

on a job knowledge test. These results suggest that the meaning of performance ratings may differ for members of different racial groups. Further research is needed on this issue.

Several other characteristics of ratees have been investigated, each in a small number of studies. Ratee age was found to have no relationship to performance ratings by Klores (1966). No ratee age effect for part-time workers was reported by A. R. Bass and Turner (1973), who did not find significant positive relationships between age and supervisory ratings for white full-time workers on one half of the dimensions being evaluated. No significant correlations between age and ratings were found for black full-time employees. Cascio and Valenzi (1977) found no effect of ratee education on supervisory ratings of police officers.

Personality factors have been investigated in a few studies. Graham and Calendo (1969) found no relationship between supervisory ratings of job performance and the ratees' personality as measured by eight scales from various personality tests. Elmore and La-Pointe (1975) reported that student ratings of college instructors' effectiveness were positively correlated with student ratings of instructor warmth.

Job-related variables. A small number of recent experimental studies have examined the effects of the performance level of the ratee on ratings of that performance. Bigoness (1976) found that actual performance had the largest effect on performance ratings. Task performance was experimentally manipulated and videotaped to standardize the stimuli for the subjects. Leventhal, Perry, and Abrami (1977) manipulated lecture quality in addition to other variables and found that student ratings of the instructor were consistently affected by the lecture quality level. Hamner et al. (1974) also found that actual performance accounted for the largest percentage of variance in performance ratings (30%), although the sex and race of the raters and ratees accounted for an additional 23% of the rating variance.

M. E. Gordon (1970, 1972) has identified what he has termed the *differential accuracy phenomenon*. He has reported that ratings were more accurate when the behavior in question was favorable rather than unfavorable. Baker and Schuck (1975) reanalyzed Gordon's data from the framework of signal detection theory and noted that the differential accuracy phenomenon appeared to be limited to only some rating dimensions and not others. The reason the effect was observed for some but not all performance dimensions was unclear, but it deserves more research attention. In a related finding Kaufman and Johnson (1974) found that negative peer nominations add little to the predictiveness of positive peer ratings. This finding is compatible with the differential accuracy phenomenon. The effect may be explainable in terms of base rates of information. Negative performance information is probably less frequent than positive information. Lay, Burron, and Jackson (1973) found that low base-rate information led to more certainty of judgment than high base-rate information. These findings, combined with those of Gordon, suggest that unfavorable information may be less accurately perceived but given more weight in the judgment process.

The effects of variability of the level of ratee performance were examined in an interesting study conducted by Scott and Hamner (1975). They manipulated the variability of subordinate performance as well as changes in the average level of performance. Variability of performance resulted in more favorable ratings of ability to do the task and less favorable ratings of task motivation but had no effect on ratings of overall task performance. A descending order of performance level led to less favorable rating of task motivation but did not affect the other two ratings. In a related study that specifically focused on the decision-making process of human judges, Brehmer (1972) found that an inconsistent cue (as variable performance could be considered) received less weight than its actual validity in a prediction task.

In a correlational field study of supervisory ratings of the performance of clerical workers, Grey and Kipnis (1976) found that the proportion of compliant and noncompliant members of a work group affected the performance ratings. A compliant worker was defined as one who had no basic job weaknesses related

to lack of ability or to an inappropriate work attitude. Ratings tended to be higher for compliant members in work groups with a large proportion of noncompliant workers than in work groups with few or no noncompliant workers. Also, ratings of noncompliant workers tended to be lower in work groups in which there were many compliant workers than in work groups with few compliant workers. The data of Grey and Kipnis, as well as those of Willingham (1958), suggest that more attention be paid to ratee group composition in research on performance rating.

Organizational and job tenure have been investigated as possible influences on performance ratings. Jay and Copes (1957) reviewed the results of 47 studies, with a total sample size of 2,462, and found that the average correlation between measures of tenure and evaluations of job performance was .17. There was a stronger relationship between tenure and performance ratings as the skill level and organizational level of the job increased. Much of the recent research in this area has supported the general findings of Jay and Copes. A. R. Bass and Turner (1973), and Cascio and Valenzi (1977), and Zedeck and Baker (1972) found positive but low correlations between tenure measures and performance ratings. Leventhal et al. (1977) manipulated the level of perceived task experience of the ratee and found that ratings of performance were higher in the condition of higher perceived experience. Some research has obtained contradictory results. Klores (1966) found no relationship between organization tenure and performance rating, although a significant positive relationship between skill level within a job family and ratings was found. Svetlik, Prien, and Barrett (1964) found a negative relationship between supervisory ratings and the job tenure of the ratee. Rothe (1949) noted that the relationship between tenure and performance ratings appeared to be affected by such factors as the organizational reward system, the intended use of the ratings, the raters' acceptance of the rating system, and the rating system's application to organizational problems. This suggestion has not been explicitly investigated to date.

Summary

The research on the effects of ratee characteristics on performance ratings offers some general conclusions. It appears that the sex stereotype of an occupation interacts with the sex of the ratee, such that males receive more favorable evaluations than do females in traditionally masculine occupations but that no differences or smaller differences in favor of females occur in traditionally feminine occupations. Ratees tend to receive higher ratings from raters of their same race, although this may not occur in highly integrated situations. Race and performance level of the ratee appeared to interact in complex ways. Further research is needed to determine if performance ratings have the same meaning for ratees of different races. Other personal characteristics of ratees have been studied too infrequently to yield conclusions about their general effects.

Experimental studies of the effect of the performance level of the ratee on performance ratings generally support the validity of the ratings. Performance level and ability have been found to have the strongest effect on ratings in these studies, although other ratee variables also significantly affect ratings. Raters may evaluate favorable performance more accurately than unfavorable, but not for all performance dimensions. Performance variability also appears to influence rating accuracy and reliability. Contrast effects may be important in performance ratings and need further investigation. Tenure and performance ratings are generally positively but weakly correlated, although situational variables may moderate this relationship.

Interaction of Rater and Ratee Characteristics

The preceding two sections have been concerned primarily with main effects of rater and ratee characteristics. This section examines the research literature that has investigated whether certain combinations of rater and ratee characteristics have effects on performance ratings.

A number of studies have been reported in which the interaction of the sex of the rater and the sex of the ratee was of interest. These

studies found no interaction effect of rater sex and ratee sex on ratings (Bartol & Butterfield, 1976; Elmore & LaPointe, 1974, 1975; Hamner et al., 1974; Jacobson & Effertz, 1974; Lee & Alvares, 1977; Rosen & Jerdee, 1973). It should be noted that the majority of these studies have involved laboratory tasks. No studies of the effects of a rater sex and ratee sex interaction on performance ratings have been reported in which both rater and ratee were actual employees of an organization. Elmore and LaPointe did investigate the ratings of college instructors by students.

The lack of rater–ratee sex interaction suggests that if the sex role stereotype hypothesis described in the Ratee Characteristics section of this article is correct, it holds for both male and female raters. Schein (1973, 1975) has reported data consistent with this interpretation. She found that both male and female managers perceived that successful middle-level managers possessed traits more commonly ascribed to men in general than to women in general. Thus, men and women seem to share common sex role stereotypes about work-related variables and could be expected to evaluate male and female ratees with common biases.

The interaction of rater race and ratee race has been investigated in several studies. The results of these studies are mixed. Schmidt and Johnson (1973) found no interaction effect of race on peer ratings in a study conducted in a highly integrated setting with individuals who had completed a human relations training program. Crooks (Note 3), DeJung and Kaplan (1962), and Hamner et al. (1974) found that raters tended to give ratees of their same race higher ratings than they gave to ratees of a different race. Crooks also found that the validity of ratings, as measured by a job knowledge test, was affected by the rater–ratee race interaction, but the results were complex. For black raters there were more valid ratings for black ratees, but for white raters nonwhite ratees received more valid ratings.

One recent study examined the hypothesis that the similarity (biographical, attitudinal, etc.) of judges and ratees affects evaluations. Frank and Hackman (1975) examined similarity effects in actual college admission interviews conducted by three college officials. They found considerable individual variation in the effect of rater–ratee similarity. One interviewer showed no similarity effects, one showed positive but weak effects, and one showed strong, positive effects of similarity. The similarity hypothesis has not been directly examined in a performance rating setting, although the data on racial similarity effects fit into this conceptual framework. Research on the similarity effect and its individual correlates appears to be a fruitful area for performance rating work.

Barrett (1966a) found that supervisor–subordinate agreement on the requirements of the subordinate's job had no effect on the mean rating or reliability of the supervisor's rating of the job performance of the ratee.

Summary

Rater sex and ratee sex do not appear to interact in their effects on evaluative judgments. Research in actual work settings is needed, however. Both male and female raters may have common sex role stereotypes that affect judgments. Raters often give more favorable ratings to same-race ratees, although situational factors may moderate this effect. It was suggested that the similarity of rater and ratee on background and attitudinal factors may affect ratings, although no direct results are available that bear on this question.

Vehicle

An enormous amount of effort has been spent exploring the potential effects of various rating formats over the years. The hypothesis has been that the vehicle that is used to elicit information has an effect on the accuracy and utility of that information. In our examination of the literature bearing on this hypothesis, we deal with methods of direct rating (in which the rater actually assigns a number to a ratee representing some level of performance), methods of derived rating (in which the rater makes a series of discrete judgments about the ratee, from which a performance rating can be derived), and technical issues, such as response categories and rating scale anchors.

Direct Rating

Graphic scales. Paterson introduced the graphic rating scale to the general psychological community in 1922. In his opinion, this new method was characterized by two things: (a) The rater was freed from quantitative judgments, and (b) the rater was able to make as fine a discrimination as desired. The scales consisted of trait labels, brief definitions of those labels, and unbroken lines with varying types and number of adjectives below. Little research was conducted on this basic format until the latter years of World War II and the postwar period. There was a growing disenchantment (Ryan, 1958) with the subjective and arbitrary nature of the graphic system. Symptoms such as leniency and halo often eliminated any potential usefulness of performance ratings. Basic research related to the properties of graphic scales was carried out by Wherry (Note 5) with armed forces personnel, but it was not widely known. In 1958, Barrett, Taylor, Parker, and Martens tested the adequacy of four different formats, varying in degree of structure. Format I consisted of a 10-inch (25.4 cm) line with 15 divisions and a trait name; there were no trait definitions or anchors on the scale. Format II consisted of the same segmented line, but trait definitions were added to the trait names. In Format III, the segmented line was defined by behavioral anchors; there were trait labels but no definitions. Format IV consisted of the segmented line defined by the behavioral anchors and trait definitions but had no trait labels. Format III showed higher reliability, lower leniency, and lower halo.

Madden and Bourdon (1964) compared several other variations of rating scale format. They varied the position of the high end of the scale, spatial orientation of the scale (horizontal vs. vertical), segmentation of the scale (segmented vs. unbroken), and numbering of scale levels (1 to 9 vs. −4 to +4). The results showed a significant main effect for the experimental manipulations taken as a whole, although the effect was small.

In a later study by Blumberg, DeSoto, and Kuethe (1966), spatial orientation of the scales was examined once again. They found no significant differences as a result of the location of the "good" end of the scale (top, bottom, left, or right). They concluded that raters might have preferences for various formats but that these preferences have little or no effect on the actual rating behavior.

The literature on logic and development of graphic scales is meager. As is demonstrated, real progress was made as a result of the introduction of alternative methods and comparisons of these methods to the traditional graphic system.

Behaviorally anchored scales.[2] In 1963, Smith and Kendall introduced a new method for rating scale development and use called *behavioral expectation scaling.* This type of scale has become alternatively known as the Behaviorally Anchored Rating Scale (BARS). Physically, it differs from traditional graphic scales in that the anchors that appear at different intervals on the scale are examples of actual behavior rather than adjectives modifying trait labels or simply numbers. More is said about the anchoring procedure in a later section. The present section deals with the use of the completed scales.

Smith and Kendall (1963) demonstrated that the scales could be used effectively for describing nursing performance. Maas (1965) demonstrated the effectiveness of the scales for measuring interview performance. Landy and Guion (1970) demonstrated the utility of the technique for ratings of work motivation. In the last 8 years, this type of rating scale has been used in a sufficient number of settings to warrant the characterization of widespread use.

In terms of direct methods of performance rating, the BARS system currently commands most attention. Nevertheless, in the course of research on this type of scale, some negative findings have emerged. There is a continuing problem with identifying anchors for the central portions of the scales (Harari & Zedeck, 1973; Landy & Guion, 1970; Smith & Kendall, 1963). There is some dispute concerning the generalizability of scales from one setting to

[2] There are several good reviews of the research in the area of developing and using behaviorally anchored scales (J. P. Campbell et al., 1973; Schwab, Heneman, & DeCotiis, 1975).

another. Borman and Vallon (1974) contended that the scales may be limited to use in the settings in which they were developed. Goodale and Burke (1975) and Landy et al. (1976) demonstrated generalizability beyond developmental settings. A similar type of scale was suggested for improving judgments in clinical settings (J. B. Taylor, Haeffele, Thompson, & O'Donoghue, 1970). Although there have been some tentative extensions of the basic BARS format (Latham & Wexley, 1977; Schwartz, 1977), the physical nature of the rating format has remained relatively constant across studies.

Almost all researchers in the area agree that a BARS is expensive to produce. The generally accepted developmental practice requires independent groups of judges (normally, samples of the population of raters who will eventually use the scales) to develop the dimensions to be rated and the definitions of these dimensions, to develop and subgroup behavioral examples of various levels of these dimensions, and finally, to assign scale values to these examples as part of the scale anchoring process. The effectiveness of the scales is assumed to be based, at least in part, on the independence of the groups engaged in each of the developmental phases. Thus, it can be seen that the investment of time is considerable. The major objection to the BARS currently is whether the ratings that these scales produce are so error free that they justify the cost of scale development. In the next section, we examine empirical comparisons of the BARS with other direct rating methods.

Comparison of graphic scales to BARS. There has been a good deal of careful work attempting to assess the relative effectiveness of the BARS in relation to traditional graphic methods of rating. J. P. Campbell et al. (1973) compared the BARS method to summated ratings and concluded that the BARS format yielded less method variance, less halo, and less leniency in ratings.

Borman and Vallon (1974) found that the BARS technique yielded ratings that were superior in reliability and rater confidence in ratings but that simpler numerical formats resulted in less leniency and better discrimination among ratees.

Burnaska and Hollmann (1974) compared three different formats. The first format was the standard behaviorally anchored scale. The second format consisted of the same dimensions and definitions, but adjectival anchors were substituted for behavioral ones. The third format was a traditional graphic rating format with a priori dimensions. Although leniency and composite halo were present in all three formats, the BARS method reduced leniency and increased the amount of variance attributable to ratee differences. Nevertheless, Burnaska and Hollmann concluded that improvements in some aspects of rating, when the BARS method was used, was accompanied by problems in other areas: "Innovations in rating, although plentiful, are likely to result in robbing Peter to pay Paul" (p. 307). Each format seemed to have its own unique problems.

Keaveny and McGann (1975) compared the student ratings of college professors on behaviorally anchored and graphic rating scales. Behaviorally anchored scales resulted in less halo, but they did not differ from graphic scales in terms of leniency. Their general conclusion was that neither format could be judged superior to the other.

Borman and Dunnette (1975) compared the standard BARS format to rating scales that had identical dimension labels and definitions but numerical anchors, and with traditional graphic rating scales that had trait labels and numerical anchors. They concluded that in spite of the fact that the standard BARS format was psychometrically superior (in terms of halo, leniency, and reliability), format differences accounted for trivial amounts of rating variance (approximately 5%).

Bernardin, Alvares, and Cranny (1976) compared summated ratings with BARS ratings. They concluded that summated ratings were characterized by less leniency and greater interrater agreement than were BARS ratings. They hypothesized that the rigor of scale development was a crucial issue in the resistance to rating errors, regardless of the format of the scales. In a follow-up study Bernardin (1977) demonstrated that when item analysis procedures are used for choosing anchors in the BARS method, there is no

difference between BARS ratings and summated ratings.

Finally, Friedman and Cornelius (1976) compared ratings from three groups: (a) a group who participated in developing BARS, (b) a group who participated in developing graphic rating scales, and (c) a group who did not participate in scale development. There was no difference in rating errors between Groups 1 and 2. The ratings of Group 3 were characterized by significantly higher levels of rating error (halo) than were the ratings of either of the other two groups.

In general, the comparisons of the BARS method with alternative graphic methods make it difficult to justify the increased time investment in the BARS development procedure. In addition, the work of both Bernardin (Bernardin, 1977; Bernardin, Alvares, & Cranny, 1976) and Friedman and Cornelius (1976) suggests that superior scales are a result of psychometric rigor in development and of some level of participation of individuals representative of those who will eventually use the scales to make ratings rather than of some characteristic unique to behavior anchors per se. More is said about the nature of anchors in a later section. In general, one must conclude that although enthusiasm greeted its introduction, the BARS method has not been supported empirically.

Derived Rating Systems

Forced-choice rating. By far the most popular alternative to direct rating schemes has been the method of forced-choice rating. In this system, the rater is required to choose from among a set of alternative descriptors (normally four items) some subset that is most characteristic of the ratee; variations of this method require the rater to choose both most and least characteristic descriptors. These descriptors function in a manner similar to that of anchors in direct rating. In direct rating schemes, the rater uses anchors to place an individual on a continuum; in a forced-choice system, the choice of descriptors by the rater allows a rating to be derived, since the descriptors have been assigned a priori scale values through some scaling process. In addition, the descriptors have usually been equated or balanced for social desirability. A review by Zavala (1965) gives a good perspective of the range of occupations and situations in which the forced-choice format has been used.

One of the assumed advantages of forced-choice rating was its resistance to leniency. This was due to the fact that the rater did not know the preference and discrimination indices of the various descriptors. Lovell and Haner (1955) found that even when students were specifically instructed to make instructors look good or bad, there was little positive or negative leniency present in the resulting ratings of their teachers. Isard (1956) found that ambiguous descriptors were more reliable and valid than either positive or negative statements and were also less prone to intentional bias. The value of neutral statements was recently confirmed in a study by Obradovic (1970) of blue-collar and white-collar performance. An attempt to use critical incidents (Flanagan, 1954) as descriptors in a forced-choice format yielded ratings with low reliabilities (Kay, 1959).

Comparison of forced-choice with other formats. As was the case with research on graphic formats, more is learned from the comparison of the forced-choice format with other formats than from an examination of variations within the forced-choice format. Staugas and McQuitty (1950) compared the forced-choice technique to both graphic ratings and peer rankings. Since the correlations between forced-choice ratings and the other two performance measures were higher than similar indices for the other two methods, it was concluded that forced-choice methodology was superior. Were they to rewrite the discussion section today, the authors would undoubtedly use the term *convergent validity* to describe advantages of the method. A number of other studies used similar logic for supporting the superiority of the forced-choice format. Berkshire and Highland (1953) demonstrated that forced-choice ratings had higher correlations with overall ranking than did graphic ratings. E. K. Taylor, Schneider, and Clay (1954) found the correlations between forced-choice ratings and graphic ratings to be high, but forced-choice ratings showed less leniency bias. Cotton and Stoltz (1960)

showed less range restriction on forced-choice ratings than on graphic ratings. It seems that these studies point to one major advantage of forced-choice ratings—they seem to maximize interindividual variance, although little is known about their effects on intraindividual variance. Since forced-choice scales were introduced primarily in an attempt to control positive bias or leniency, little attention was paid to the problem of halo error. Nevertheless, there is some peripheral evidence available. Sharon and Bartlett (1969) examined the relative resistance of forced-choice and graphic ratings to leniency bias under four conditions: (a) rater anonymous, research purposes only; (b) rater anonymous, feedback to instructor; (c) rater identified, research purposes only; and (d) rater identified, follow-up discussion with ratee. The ratings represented student evaluations of college instructors. Although the results showed significant leniency for all graphic rating conditions, the forced-choice ratings were uniformly resistant to leniency bias.

In one of the few comparative studies dealing with reliability, Lepkowski (1963) found that graphic ratings and forced-choice ratings yielded equal reliability, when scales were developed from critical incidents; from the earlier studies of descriptors reported previously, it might be concluded that these descriptors worked to the disadvantage of the forced-choice format.

E. K. Taylor and Wherry (1951) compared forced-choice and graphic rating systems in a military setting. In one condition, they told raters that the ratings were for experimental purposes; in a second condition, they implied that the ratings would be used to make administrative decisions. They found an increase in the mean ratings for both formats under the "for keeps" condition. Nevertheless, the impact seemed to be greater on the graphic ratings. In addition, there was poorer discrimination among ratees at the top of the scale in the graphic, for keeps condition. They suggested that graphic rating scales that followed forced-choice scales might be less biased.

In 1959, Cozan reviewed the studies that addressed the validity of the forced-choice format and concluded that unless a new system was clearly superior to an existing system, the costs of organizational change argued in favor of the old system. Since the studies to that date had not presented any compelling reason for choosing forced-choice formats over alternative formats, he suggested that traditional graphic rating schemes be retained. The argument is similar to that made against the increased cost of developing BARS formats. Nevertheless, there does seem to be evidence that the forced-choice format reduces range restriction. Unfortunately, sufficient data are not available to determine what price is paid for this psychometric advantage.

Recently, a variation of the derived rating format has appeared. Blanz and Ghiselli (1972) suggest a method in which the rater is required to indicate if the ratee is better than, equal to, or worse than the behavior presented. Since these behaviors have been previously scaled in terms of the level of performance that they represent, it is possible to derive a rating from these judgments. Since behaviors from many different dimensions are randomly arranged, it is difficult for the rater to determine order of merit values for the various statements or to determine which dimensions are being measured. This is thought to protect against intentional bias. Although it is much too early to draw any firm conclusions about this technique, some early results (Saal & Landy, 1977) have been disappointing. Although halo errors are lower in this format than in graphic or BARS formats, reliabilities seem to be exceptionally low. In addition, there are some serious problems with the scoring formats suggested in the original presentation of the method. Arvey and Hoyle (1974) found that scales developed by Guttman scaling techniques (the methodological foundation of the mixed standard scale) demonstrated good convergent and discriminant validity but that attempts to use the technique to identify poor raters were not successful. There was only slight evidence to suggest that raters who made rating errors on one job dimension also rated poorly on other dimensions or that raters who made errors in rating one individual also made errors when rating other individuals. The Guttman-based scales also exhibited higher rating inter-

correlations than more traditional behavioral-based scales.

Rating Dimensions

The nature of what dimensions are to be rated has created some controversy. Kavanagh (1971) argued that the empirical literature does not allow one to choose unequivocally the type of content of rating scales (i.e., performance results vs. observable job behaviors vs. personal traits of the incumbent). Brumback (1972) has called for the elimination of personal traits as rating dimensions in preference to performance factors. Kavanagh (1973) rejoined that the question of appropriate rating dimensions can only be answered by a consideration of the nature of the job requirements, the relevance of personal and performance factors for that job, and the empirical defensibility of each content type in terms of reliability, resistance to rating errors, and construct validity.

Several authors (e.g., J. P. Campbell et al., 1970; James, 1973; Ronan & Prien, 1966; Smith, 1976) have argued that global ratings of job performance are more likely than more specific ratings to be affected by extraneous sources of variance and to omit important relevant sources of variance.

Number of Response Categories

A series of careful studies by Bendig (1952a, 1952b, 1953, 1954a, 1954b) provide firm evidence concerning the most efficient number of response categories for rating formats. Considering both scale reliability and rater reliability, there is no gain in efficiency when the number of categories increases from 5 to 9; reliability drops with 3 categories or less and with 11 categories or more. Finn (1972) studied the effect of number of categories on reliability of rating and found that reliability dropped with less than 3 or more than 7 response categories. Lissitz and Green (1975), in a Monte Carlo study of the effect of response categories on scale reliability, concluded that there was little increase in reliability when there were more than 5 scale points or response categories. Bernardin, LaShells, Smith, and Alvares (1976) compared a continuous to noncontinuous 7-point response format and found no differences in rating errors. Finally, Jenkins and Taber (1977), in another Monte Carlo study of factors affecting scale reliability, agreed with Lissitz and Green that there is little utility in adding scale categories beyond 5.

Since several studies have found that an excessive number of categories can have negative effects on scale reliability, the studies on number of response categories seem to have serious implications for scales that allow the rater to define the number of categories of response. Specifically, the format should require the individual to use one of a limited number of response categories, probably less than nine. In spite of the fact that the Bernardin, LaShells, Smith, and Alvares (1976) results do not support the generalization, the weight of evidence suggests that individuals have limited capacities for dealing with simultaneous categories of heterogeneous information. This was suggested long ago by Miller (1956) in his now-famous seven, plus or minus two dictum and appears to generalize to rating behavior.

Anchors

Type. Rating scales typically have one of three types of anchors: numerical, adjectival, or behavioral. There have been several studies directed toward determining the relative effectiveness of these alternative anchoring systems. This, of course, begs the question of whether there should be any anchors at all. Several studies have demonstrated the positive effect of increasing the degree of scale anchoring. Bendig (1952a, 1952b, 1953) found that scale reliability improved with increased anchoring. Barrett et al. (1958) demonstrated the increased effectiveness of anchored scales compared to unanchored ones. D. T. Campbell, Hunt, and Lewis (1958) examined the effect of shifting context on ratings of cognitive organization that are present in responses of schizophrenics. They found that a nine-point scale with detailed anchors was less susceptible to distortion than a similar scale with a minimum of descriptive anchoring.

A number of studies have suggested the

relative effectiveness of behavioral anchors as compared with simple numerical or adjectival anchors (Barrett et al., 1958; Bendig, 1952a, 1952b; Maas, 1965; Peters & McCormick, 1966; Smith & Kendall, 1963). Since the BARS format has relied heavily on the behavioral nature of the scale anchors, almost all studies positive toward the BARS format might also be thought of as positive toward behavioral rather than adjectival or numerical anchors. Nevertheless, there have been some studies that have cast some doubt on the nature of elaborate anchors. Finn (1972) found no difference in means or reliabilities of ratings as a function of the manner of defining scale levels. It did not seem to matter if the anchors were numerical or descriptive. Kay (1959) found that using critical incidents to anchor rating scales depressed reliability values. He suggested that critical incidents were too specific and context bound for use as anchors.

The importance of the type and number of anchors probably covaries with the adequacy of the dimension definition. In the absence of adequate definitions of the dimensions to be rated, the rater must depend on the anchors to supply the meaning of the scale. In the Barrett et al. (1958) study, it was found that scales with good behavioral anchors and no dimension definitions (only trait labels) had higher reliability, less halo, and less leniency than either scales with definitions and anchors or scales with definitions but no anchors. In general, it seems that anchors are important, and there is some evidence to suggest that behavioral anchors are better than numerical or adjectival ones.

Scaling. Traditionally, one of the weaker procedures in the development of rating scales has been the process of assigning scale values to anchors. Although there have been some attempts at modifying scale intervals with numerical and adjectival anchors (Bendig, 1952a, 1952b), for the most part there has been little research on the nature of the psychological scale of measurement implied by traditional graphic rating scales.

In the forced-choice format, indices of discrimination and favorability are determined for each item to be used. The discrimination index is the degree to which the particular item or phrase discriminates between high and low performers; the preference value is the degree to which the trait or behavior represented by the item is valued by the typical rater. Both Isard (1956) and Obradovic (1970) found that neutral items were superior to positive or negative items in terms of psychometric characteristics of the resulting ratings. By implication, one might conclude that items with intermediate scale values on the preference dimension were more useful than high- or low-preference items.

Smith and Kendall (1963) used item analysis to determine the ability of particular behavioral anchors to discriminate good from poor nurses. Unfortunately, few studies that followed the Smith and Kendall lead were rigorous in terms of either selection or placement. The most common technique for anchoring in the BARS method is the use of judges to estimate how much of a particular dimension a behavioral example represents. This is usually accomplished with graphic ratings on adjectivally or numerically anchored graphic rating scales. These items have been previously judged for content. The final selection of anchors is based on the resulting mean value of the item and on its standard deviation. The decision rule is frequently arbitrary. (Choose items that represent as many points along the continuum as possible, and have standard deviations less than a specified value.) Bernardin (Bernardin, 1977; Bernardin, Alvares, & Cranny, 1976) demonstrated that the resistance of rating scales to traditional rating errors depends, to some degree, on the rigor of scale development and anchoring. He suggests that this rigor can be introduced through standard item analysis procedures. Barnes (Note 6) demonstrated that a Thurstone solution of a pair comparison scaling of behavioral anchors produces anchor scale values that are significantly different from those produced by the graphic rating method common to BARS development. It may very well be that the reason for the relatively disappointing showing of the BARS format compared with other formats has been due to a lack of rigor in the selection and scaling of anchors. This has been suggested by Schwab, Heneman, and De-

Cotiis (1975), Bernardin (1977), and Landy (Note 7).

There are some data which suggest that the anchoring process itself is susceptible to the same kinds of biases that new rating formats, such as BARS, are attempting to eliminate. Thus, the situation becomes one of infinite regress; the anchoring procedure introduces the very type of error variance that the format attempts to eliminate. Wells and Smith (1960), Rotter and Tinkleman (1970), Landy and Guion (1970), and Barnes (Note 6) have all demonstrated that anchoring procedures affect ultimate item scale values and standard deviations. These data suggest that anchoring be accomplished by means of some method based on firm psychometric theory (such as Thurstone's law of comparative judgment).

Summary

After more than 30 years of serious research, it seems that little progress has been made in developing an efficient and psychometrically sound alternative to the traditional graphic rating scale. Nevertheless, we have learned some things about rating formats in general. In spite of the fact that people may have preferences for various physical arrangements of high and low anchors, of graphic numbering systems, and so forth, these preferences seem to have little effect on actual rating behavior. The number of response categories available to the rater should not exceed nine. If a continuous rather than discrete response continuum is contemplated, it would be wise to conduct some pilot studies to determine how many response categories are perceived by the potential raters. There is some advantage to using behavioral anchors rather than simple numerical or adjectival anchors. This advantage is probably increased in the absence of good dimension definitions. Finally, it is important that rigorous item selection and anchoring procedures be used in the development of rating scales, regardless of the particular format being considered. It may be that new techniques, such as the BARS or the mixed standard, will show improvements over more traditional methods, if more rigorous developmental procedures are used.

Context

Included in this section are studies that have examined the effects of factors that are not explicitly related to the nature of the rater, ratee, or rating instrument but that may be considered as part of the context in which the rating occurs.

A number of studies have investigated the effect of the intended use of the ratings on various psychometric properties of the ratings. Several studies have shown that ratings are more lenient under conditions of administrative use than under conditions of research use (Borresen, 1967; Heron, 1956; E. K. Taylor & Wherry, 1951; Centra, Note 4), whereas Sharon (1970) and Sharon and Bartlett (1969) found a similar effect for graphic rating scales but not for a forced-choice scale. Bernardin (1978) found that leniency decreased in conditions in which the importance of the ratings was stressed. Kirkpatrick et al. (1968) reported that whereas ratings on three scales intended for research purposes only did not differ for black and white ratees, ratings for the same sample of ratees on a scale intended for administrative use were more favorable for the white employees. Hollander (1957, 1965) found no difference between administrative and research conditions in terms of the reliability or validity of ratings.

A few investigations of the effect of position or job characteristics on performance ratings have been conducted. Much of this literature has been concerned with the sex role stereotype hypothesis discussed in previous sections of this review. In general, the data suggest that ratings are influenced by the interaction of the sex of the ratee and the sex role stereotype of the job or task, although not all studies support this general conclusion (e.g., Jacobson & Effertz, 1974).

In other research that examined position characteristics, Svetlik et al. (1964) found that job difficulty, as measured by a job evaluation point system, was weakly but positively related to a supervisory rating of job competence, although not to a rating of over-

all effectiveness. Klores (1966) reported that performance ratings were positively correlated with skill levels within a job classification. Myers (1965) partialed job level from rating intercorrelations and found a reduction of halo effects in the correlation matrix and a more meaningful factor structure.

Several other contextual effects on ratings have been reported. Cascio and Valenzi (1977) hypothesized that the expectation of favorable performance because the ratees had passed selection and training hurdles might cause raters to be lenient in their judgments. Rosen and Jerdee (1973) found that observer ratings of the effectiveness of supervisors' leadership styles were affected by the sex of the supervisor and the sex of the subordinates. In general, ratings were more favorable when the leadership style of the supervisor was appropriate for traditional sex role stereotypes. Rothe (1949) found that the nature of the incentive system could affect ratings. When the ratees were given pay increases based on performance ratings, and the pay for each job had a ceiling, supervisors gave more favorable ratings to less senior subordinates who had not yet reached the pay maximum.

Summary

Ratings for administrative purposes will be more lenient than those for research purposes. Unfortunately, since most of the published research was done in the research purposes context, too little information is currently available to draw firm conclusions about impact of purpose for rating. Although it does not appear that the variances of the ratings are affected by the purpose component, more definitive tests of this relationship are needed.

Rating Process Variables

Another class of variables affecting performance ratings are those factors related to the process by which ratings are obtained, exclusive of the rating instrument itself. Included here are such variables as rater training, rater anonymity, and sequence of traits and ratees.

A large number of primarily recent studies have examined the effect of rater training on rating errors and validity, following the suggestions for such research from Borman and Dunnette (1975), Moore and Lee (1974), and Schneier (1977), among others. Most investigations have found that training raters reduces rating errors, although some data suggest no differences between trained and untrained raters (J. B. Taylor et al., 1970; Vance, Kuhnert, & Farr, 1978) or only short-term effects (Bernardin, 1978). Wexley, Sanders, and Yukl (1973) found that only extensive training was effective in reducing rating errors, a finding corroborated by Brown (1968), Latham, Wexley, and Pursell (1975), and Bernardin and Walter (1977). Borman (1975) reported decreased halo, with no reduction in the validity of ratings, with only a brief training program.

Although rater training programs have concentrated on the avoidance of the typical rating errors such as halo and leniency, only Bernardin (1978) has demonstrated a correlation between knowledge of such errors, as measured by a test, and a reduction of the errors in actual ratings. Data from a few other studies suggest that the nature of rater training programs should include instruction in more than rating errors. M. E. Gordon (1970) found that greater experience with a particular rating instrument improved rater accuracy, and Friedman and Cornelius (1976) found that rater participation in rating scale development resulted in decreased rating errors, regardless of scale format. These studies can be interpreted as pointing to the inclusion of detailed instruction in the use of whatever scale format has been selected. Klieger and Mosel (1953) noted that better interrater agreement among supervisory raters, when compared to peer raters, might be the result of supervisory training that gave raters a common frame of reference for the evaluation of performance. Rater training should also stress performance requirements of the job as well as instruction in observation techniques.

Bayroff et al. (1954), Sharon and Bartlett (1969), and Stone, Rabinowitz, and Spool (1977) found no difference in rating errors or validity between identified and anonymous raters. Creswell (1963) reported no effect on

leniency of ratings but did find more variance with confidential ratings as opposed to ratings that were to be shown to the ratees or to the rater's superior.

A suggestion for the reduction of halo error by Guilford (1954), among others, was to rate all ratees on a given trait or dimension, then to rate all ratees on the next trait, and so forth. This was predicted to result in less halo error than the process of rating a given ratee on all traits, then rating the next ratee on all traits, and so forth. However, studies that have compared these two rating processes have found no differences (Blumberg et al., 1966; Brown, 1968; Johnson, 1963; E. K. Taylor & Hastman, 1956). Johnson (1963) is a reanalysis of Johnson and Vidulich (1956), using more appropriate statistical techniques that resulted in the conclusion by Johnson (1963) of no effect.

Several studies have reported data concerning the question of whether position of the ratee in a sequence affects the rating of the ratee. Bayroff et al. (1954) found that ratings early in a sequence were more valid than those later in the sequence. Wagner and Hoover (1974) reported that raters who were not especially knowledgeable about the technical aspects of the task performance being evaluated tended to be more favorable to ratees early in the sequence. Finally, Willingham (1958) found that ratings tended to be biased in the direction of the previous rating and that this tendency increased as the number of response categories increased.

In an isolated study of a rating process variable, Wright (1974) found that when faced with less time to reach a judgment, individuals tended to use fewer sources of information and to weigh unfavorable information more heavily in making evaluations.

Summary

Rater training has generally been shown to be effective in reducing rating errors, especially if the training is extensive and allows for rater practice. Questions still remain about the longitudinal effects of such training, about the effect of rater training on the validity of ratings, and about the optimal content of such training programs.

Identified raters, as opposed to anonymous raters, appear to give equivalent ratings. There appears to be no reduction in halo error when all ratees are evaluated on one trait, then all ratees are evaluated on the next trait, and so forth. Serial position appears to have some effect on ratings, but no general pattern has emerged from the research to date.

Results of Rating

After one gathers ratings of performance, decisions must still be made concerning the manner in which these data might be analyzed to produce accurate and reliable performance descriptions. It is possible that various analytic techniques are more successful at reducing or eliminating rating errors than other techniques. In this section, we review the research that addresses this issue.

Dimension Reduction

Traditionally, the performance of individuals is considered with respect to a number of presumably independent dimensions. We use the term *presumably* because, in spite of the attempt by the research to identify and define independent aspects of performance, the intercorrelations among ratings on these dimensions are often high; this problem is usually introduced as one of halo error. A number of studies have examined the effect of combining ratings across dimensions into some smaller number of homogeneous subsets. The process of combination is assumed to produce new, derived performance scores that are more reliable and that better describe important aspects of work-related performance, allowing the construction of more efficient selection and training programs.

The most common technique for data reduction and combination in the area of performance ratings has been factor analysis. Factor scores have been computed and used as the criteria for various administrative, counseling, and research purposes. Grant (1955) suggested that factor analysis was a useful device in determining the degree of halo present in ratings, as well as the degree to which discriminations might be made among individuals on performance dimensions.

Guilford, Christenson, Taaffe, and Wilson (1962) proposed that certain marker tests and variables be included in a factor analysis of performance ratings to better understand exactly what was being rated. Schultz and Siegel (1964) gathered estimates of judged similarity among performance dimensions from raters and used those judgments as the basis for identifying more basic performance dimensions through multidimensional scaling procedures. Dickinson and Tice (1977) used factor analytic techniques to improve the discriminant validity in performance ratings. They suggested that factor analysis could be used to eliminate complex anchors (anchors loading on more than one dimension) and thus reduce trait intercorrelations. Kane and Lawler (1978) suggested that performance ratings be factor analyzed and that the first unrotated factor score be used as a measure of overall performance for administrative purposes.

Unfortunately, there have been no rigorous tests of the hypothesis that reduced scores tell us more about the performance of the ratee than do raw scores. As a matter of fact, there are several studies that have implied that a factor analysis of ratings tells us more about the cognitive structure of the raters than about the behavior patterns of the ratees. Grant (1955) suggested that factor analysis might tell us how the raters interpreted the items on the rating scale. Norman and Goldberg (1966) instructed raters to evaluate individuals with whom they had little familiarity and individuals with whom they were quite familiar. The factor structures derived from a factor of the two data sets independently were similar. They concluded that the similarity among these dimensions was more a property of the raters' views of performance than of the ratees' actual behaviors. This has serious implications for the use of factor analytic results for developing selection or training programs. Kavanagh, MacKinney, and Wolins (1971) demonstrated the difficulties in determining the number of performance dimensions represented in managerial ratings. Using multitrait–multimethod procedures (D. T. Campbell & Fiske, 1959), they had serious difficulties in identifying independent aspects of managerial performance. They were able to identify "personality dimensions" suitable for rating. They imply that for the first time, a procedure was available for identifying and measuring personality-based performance aspects in managers. This begs the question of the existence of these dimensions in the behavior patterns of the ratees. It may be that they exist in the cognitive framework of the raters and have little or no reality with respect to ratees.

In general, this particular area of research raises many more questions than it answers. The identification of behavioral patterns in ratees is an extremely complex issue. Data reduction techniques such as factor analysis and cluster analysis of ratings seriously confound dimensions of rater cognitive structure with ratee behavior. More is said about this problem in a later section on implicit personality theory.

Weighting and Grouping Methods

Factor analysis might be thought of as one technique for assigning weights to performance dimensions; these weights would represent relative importance and could be calculated on the basis of variance accounted for by a particular component or factor. Jurgensen (1955) suggested that statistical weights might help improve the relationship between performance ratings on specific dimensions and overall performance. He found that statistical weights were no better than arbitrarily assigned weights. In light of the recent research of Dawes and Corrigan (1974), this finding might be extended to include unit weights. Naylor and Wherry (1965; Wherry & Naylor, 1966) were able to describe distinctly different rater "policies" in rating behavior. Each of these different policies could be described by a unique set of relative weights. Consequently, if different raters have different policies, it is likely that computing statistical weights on a group basis will not be effective in weighting specific dimensions—some initial subgrouping is required. Passini and Norman (1969) suggested a complicated weighting scheme for improving the reliability of peer nominations and rankings that might be extended to cover ratings as well. They suggest that indices of agreement be com-

puted within and across ratees; dimensions with greater agreement across raters would receive heavier weights, and dimensions that have highest interobserver agreement within ratees would also receive heavier weights.

The Passini and Norman (1969) procedure assumes that two or more raters are available for each ratee. Several researchers have suggested that multiple ratings are desirable. Carter (1952) suggested that multiple ratings would improve criterion reliability. Windle and Dingman (1960) found that perceptions shared by two raters were more predictive of an independent criterion than were perceptions unique to each judge. This study was in response to an earlier study by Buckner (1959) that proposed that unique views of individual judges were ultimately more valuable than common views of a ratee. Overall (1965) suggested an optimal weighting scheme based on the reliability and variance of individual raters. He was concerned with methods for combining ratings from multiple judges. He developed a scheme for weighting judgments by factors that represented the reliability and variance of individual raters. Einhorn (1972) proposed combining components of expert judgment to take advantage of differential validity of individual judges. All of these suggestions assume the existence of multiple ratings on a single ratee. This is seldom the case in applied settings. Nevertheless, even if it were possible to accumulate multiple ratings, the results of Naylor and Wherry (1965; Wherry & Naylor, 1966) and Passini and Norman (1969) imply that there are significant differences among raters in rating policies. This, in turn, implies that gains produced by collapsing ratings across judges may be illusory.

There have been some suggestions that scoring schemes might be developed for improving various types of rating data. Meyer (1951) suggested a binary scoring scheme based on the absolute number of items checked in a checklist type rating scale. He found no differences between this simple technique and more complicated ones. B. M. Bass (1956) suggested a binary scoring system that was based on the relative distribution of superior and inferior performers on a five-point scale. The simple binary recoding system reduced leniency without significantly reducing internal consistency. There have been some recent studies in scoring procedures, but they have been confounded with scale format and instructions (Bernardin, LaShells, Smith, & Alvares, 1976; Blanz & Ghiselli, 1972; Zedeck, Kafry, & Jacobs, 1976), and the results do not easily generalize to weighting and grouping principles.

Other Techniques of Statistical Control

Myers (1965) attempted to remove halo from the ratings of job factors in a job analysis study of 82 different jobs on 17 dimensions. He partialed out the effect of organizational level on these ratings and substantially reduced dimension intercorrelations. As a result, he suggested that researchers look for demographic factors that correlate highly with ratings and that remove the influence of those factors statistically, thereby reducing unwanted halo.

Ritti (1964) substantially reduced halo in supervisory ratings by standardizing both the rows and the columns of the rating matrix. He was able to demonstrate through factor analytic results that the simple structure was improved, that the within-factor correlations remained high, and that the between-factors correlations were substantially reduced.

Landy and Vance (Note 8) partialed out an overall rating from ratings on 15 specific performance dimensions and found a substantial reduction in factor correlation, a reduction in variance accounted for by the first unrotated factor (a possible measure of method variance) and an improved simple-structure factor solution.

Summary

The research on dimensional combination strategies resulting from data reduction algorithms has been equivocal. A substantial amount of research still must be done before it will be possible to specify what those performance factors represent—behavioral patterns of ratees or cognitive constructs of raters. The research on the combination of multiple ratings is also equivocal. If distinct patterns of rating exist among raters (i.e.,

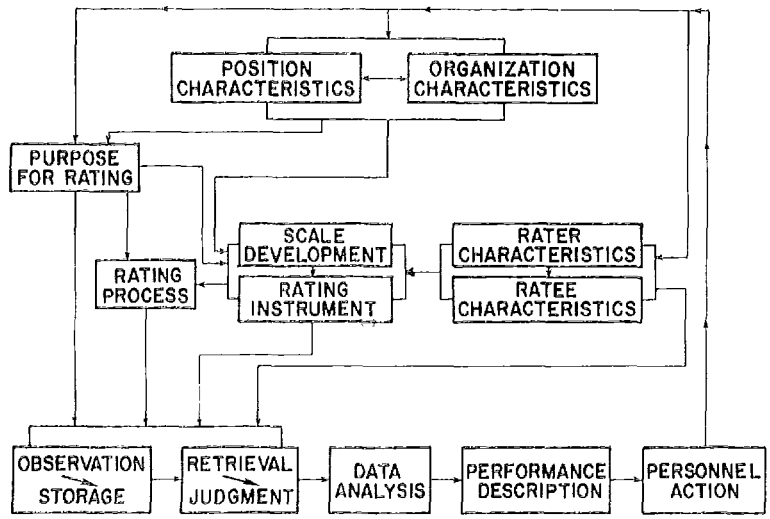

Figure 2. Process model of performance rating.

policies), the presumed gains in reliability may represent a hollow victory. On a practical basis, it seems unlikely that techniques based on multiple ratings will prove useful in applied settings, although they may be of value in cataloging sources of error. The research by Myers (1965), Ritti (1964), and Landy and Vance (Note 8) suggests a possible avenue for continued research in data analysis procedures.

Process Model of Performance Rating

Many researchers in the area of performance rating have concluded that a model of some sort is necessary before any significant advances can be made in understanding judgmental performance measures (DeCotiis, 1977; Jenkins & Taber, 1977; Kane & Lawler, 1978; Schwab et al., 1975; Zedeck, Jacobs, & Kafry, 1976). Figure 1 was presented earlier as a first or primitive representation of the rating process. On the basis of our review, as well as on some theoretical influences that are described below, we propose that Figure 2 is a more refined, coherent, and catholic representation of the system of performance rating. In Figure 2, we have tried to be more specific about the subsystems that form the larger rating system. Thus, the context component of Figure 1 has been broken into three components: position characteristics, organization characteristics, and the purpose for rating. Similarly, the rating process component now comprises two subsystems: the cognitive process of the rater (observation, storage, recall, and judgment) and the administrative rating process of the organization. With this general evolution in mind, we now discuss specific components of the model.

The model assumes that there are certain characteristics brought to the rating task that are properties of the rater and ratee, respectively. For example, the rater brings to the task "sets" or biases that may be related to age, sex, race, leadership style, personal relationship to ratee, and so forth. In addition, the ratee possesses certain characteristics, in addition to level of performance on the dimension under consideration, that may influence the judgment. In addition to the main effects that these respective rater and ratee characteristics represent, there are undoubtedly interactions of rater and ratee characteristics. Thus, in addition to the fact that black ratees may receive lower ratings than white ratees and in addition to the fact that black raters may be harsher in rating than white raters, there is also the possibility of an interaction of the race of the rater and the race of the ratee in performance judgment. The same would be true of other rater–ratee characteristics. The model also implies that the characteristics of the rater and ratee have an

influence on the selection and/or development of the rating instrument. These characteristics might include education, previous experience with performance rating, and tenure in the organization.

The most immediate context in which rating occurs is defined by the particular organization and the particular position under consideration. Organizational size moderates such critical variables as span of control, seasonal variation in work force, levels of turnover, part-time to full-time employee ratio, and so forth. It has also been proposed that organizations differ with respect to the climate that is perceived by their members. In addition to differences among organizations, there are distinct differences among positions within those organizations. Positions differ with respect to level within the organizational hierarchy; they also differ with respect to line versus staff and blue collar versus white collar designations. We propose that position and organizational characteristics jointly affect both the choice and/or development of a rating instrument and the purpose for which rating is done. It is not uncommon to see ratings used to make administrative decisions at one level in an organization but used for counseling at another level. In addition, supportive organizations often use ratings for employee development, whereas punitive organizations might use the same information for employee terminations. It is appalling to note how little systematic research has addressed the impact of position and organizational characteristics on performance rating.

The purpose component of the model is of central importance.[3] Employee counseling usually requires an instrument different from that to be used simply for administrative purposes. In addition, the purpose for the rating also affects the rating process. As an example, employees are often permitted to examine supervisory comments in development or training contexts, but they are usually not permitted similar access in the context of salary decisions. Finally, the purpose for ratings is assumed to have a substantial effect on the cognitive process of the rater. As indicated in the earlier review, ratings done for research purposes differ substantially from those done for administrative purposes.

A conceptually independent variable in the system is the vehicle or instrument actually used to gather the performance information. Through a process of scale development or selection, an instrument is identified that presumably is capable of helping raters make distinctions among ratees with respect to various categories of behavior. The scale development may involve developmental groups, as in the case of the BARS methodology, or item analysis derived from a study of current employees, as in the case of summated ratings or forced-choice inventories. Regardless of the method of development, an instrument will be selected or constructed to produce judgments about performance. This instrument will have certain phenotypic characteristics—type and number of anchors, number of response categories, spatial orientation, and so forth. As indicated previously, these characteristics will be directly influenced by the rating context (including position characteristics, organizational characteristics, and purpose for rating) and the characteristics of the rater and ratee. The final instrument will have a major impact on the cognitive operations of the rater.

The component labeled rating process refers to the constraints placed on the rater by requests or demands. For example, ratings gathered once a year will have different characteristics than those gathered twice a year; ratings gathered on the same day for all ratees will have different characteristics than those gathered on the anniversary of the date of hire for each ratee; ratings done in a noisy, public, distracting environment will have different characteristics than those done in a quiet, private, distraction-free environment; ratings that will eventually be seen by the ratee will have different characteristics than those that will not be seen by the ratee; rating sessions preceded by brief training modules will yield ratings that have different characteristics than those produced "cold," and so forth. As indicated earlier, the rating process is not developed in isolation. It will inevitably be influenced by the purpose for rating and the instrument used for the rating.

[3] We are grateful to Shelly Zedeck for suggesting that this is an explicit component of the model rather than a derivative component.

The cognitive operations of the rater are thought to fall into two temporal categories. In the first category we find observation and storage. Usually, the rater is asked to retrieve and use that information at some later time. These cognitive operations are influenced by a multitude of variables: the purpose for the rating, the administration rating process, the instrument used, and interacting rater–ratee characteristics.

After performance judgments are made by the rater, a decision must be made concerning how to treat that information. As we have seen in the literature review, these decisions may have substantial effects on error variance estimates. One might use influences identified in previously discussed rater, ratee, position, or organizational components as covariates or moderators. The way in which the data are treated must be considered as a potential source of variance in the resulting performance description. The long-standing argument regarding composite versus multiple criteria is evidence of the importance of this component in a general performance measurement system (Landy & Trumbo, in press). If information is improperly combined and inaccurately fed into the personnel decision system, it will inevitably have negative effects on that system.

The component labeled performance description implies that the data analytic procedure yields a result that is fed back to appropriate sources. These sources might be personnel departments engaged in validation studies or training evaluation studies. They might include administrative officers engaged in salary or work force planning for the coming year. They might include ratees who receive performance feedback interviews from supervisors. In a sense, each of these sources engages in a personnel action either actively or by default—selection systems are maintained or changed, salaries or work force levels are maintained or changed, employees are told of strengths and weakness or ignored. These actions, or lack thereof, influence the characteristics of both raters and ratees. If layoffs result from rating data, both raters and ratees will be changed by virtue of that action. Rater biases or sets and ratee behavior will most likely both change. Presumably, accurate performance data fed back in a nonpunishing manner will help individuals eliminate weakness and maintain strengths.

The feedback loop from personnel action to purpose for rating has an important implication. It implies that we must consider how the purpose for rating is perceived by the rater; it is the resulting personnel action that will clearly inform the rater of the purpose for rating rather than the organization's stated purpose, if the two are at odds.

As can be seen from the descriptions of the model in Figure 2, there is a good deal of research still to be done if we are to understand the nature of performance ratings in any nontrivial way. The model is strongly process oriented and must be eventually supported by more substantive propositions concerning where rater biases come from, why certain rating processes minimize rating error, whereas other procedures exaggerate those errors, and so forth. Two substantive approaches that may be useful to use in conjunction with the process model previously described are implicit personality theory and Wherry's (Note 5) psychometric theory of rating. We now describe how these approaches complement and support our process model.

Implicit Personality Theory

For a substantive theoretical framework to be of value in understanding a behavioral phenomenon, such as rating of performance, it should be able to unify the various manifestations of the phenomenon; it should have the capacity to explain conflicting results, to bring order to disorder. The major theme in the research that has been conducted in the area of performance rating has been that variables of major importance can be found in the rating scales themselves. Individual differences in raters were only occasionally investigated. Even when these differences were examined, they tended to be second-level or demographic differences, such as sex or experience, rather than first-level direct influences, such as cognitive operations or feelings toward the stimulus object.

It is our feeling that implicit personality theory research is rich with implications for understanding rating behavior. Bruner and

Tagiuri (1954) considered implicit personality theory to be assumed relationships among traits. This is close to the definition of halo in performance rating that one might infer from reading the literature that we have reviewed in the earlier sections of this article. Cronbach (1955) expanded this definition to include not only the covariation among traits but also the means and variances of the traits, implying some relationship between implicit personality theory and leniency and central tendency errors.

In some senses, performance ratings may represent specific instances of implicit personality theories of raters—assumed values on performance dimensions that are independent of actual behavior of the ratee on those dimensions. The work of Passini and Norman (1969) and Norman and Goldberg (1966) is interesting in that respect. They were able to demonstrate that the correlation among traits could be accounted for through constructs of the rater rather than through co-occurrences of behavior patterns in ratees.

An earlier study by Koltuv (1962) demonstrated that implicit personality theories were most likely to operate in situations in which there was low familiarity of rater with ratee. Thus, one might reasonably conclude that the means, variances, and covariances of performance ratings will depend to a certain extent on the degree of familiarity of rater with ratee. This will not come as a surprsie to anyone who has conducted research in the performance rating area; nevertheless, it represents a more general phenomenon, since it has been found in settings other than that of performance ratings. In addition, the finding takes on more cognitive overtones when it is embedded in other findings in the implicit personality theory research. For example, implicit personality theories seem to operate more often in trait evaluations than in behavioral evaluations (D. J. Schneider, 1973). Particularly, trait labels seem to lead more often to implicit assumptions. Together, these findings suggest that rating scales should be behaviorally anchored with no trait labels at all. It may be that the labels that introduce the scales create response sets that the definitions and anchors are unable to eliminate. It would be easy enough to test this hypothesis by examining the intercorrelations of ratings made with labels and without labels; although the presence or absence of definitions has been examined, and the presence or absence of behavioral anchors has been examined, little attention has been paid to dimension labels, which invariably read as trait names.

There has been also the suggestion in implicit personality theory research that we seek less information about persons we dislike or about individuals from different strata. This is similar to the Differential Accuracy Phenomenon (DAP) described by M. E. Gordon (1970, 1972) that led him to conclude that low performers are described less accurately than high performers; in addition, the number of levels in the organization that separate the rater and ratee seem to affect the accuracy of the performance judgment (Whitla & Tirrell, 1953; Zedeck & Baker, 1972).

The interesting aspect of implicit personality theory as a heuristic device for examining performance ratings is that it brings a new outlook to some old and vexing questions. It requires us to view rating errors in a new light. These errors become behavioral phenomena governed by individual differences. Landy et al. (1976) suggested that rating errors are not simply properties of scales or instruments. Errors are governed by many parameters, and some of these parameters may be cognitive differences among raters. One might apply Kelly's (1955) notions of personal constructs to performance rating and come to the same conclusion. As another example, Tajfel and Wilkes (1963) found that subjects made more extreme judgments on dimensions that they provided themselves rather than on those given to them. What would happen if we were to allow raters to choose a subset of dimensions from a larger number of dimensions? Would halo be reduced? Would central tendency be reduced? Would new errors appear? The point we are trying to make is that we have a substantive body of research more clearly tied to interpersonal evaluation than currently exists in the performance evaluation literature. This body of research appears under the rubric of implicit personality theory. We feel that advances can be made in understanding rating

behavior, if we view rating as a specific instance of the more general phenomenon of person perception.

Psychometric Theory of Rating

Wherry (Note 5) in 1952 proposed a theory of the rating process that has received little attention. He noted that rating accuracy is dependent on (a) the performance of the ratee, (b) the observation or perception of the ratee's performance by the rater, and (c) the recall of the observation of performance by the rater. Each of these three components of a rating may be broken into subcomponents. Drawing on classical mental test theory (e.g., Gulliksen, 1950), Wherry assumed that each component of the rating process had a systematic part and a random part. The systematic part can be further partitioned into a true aspect and a bias aspect, where bias is systematic.

In particular, the actual job performance of the ratee at a given time is equal to the sum of the true level of the ratee's job performance, environmental influences on the ratee's performance (bias), and random error. Each of these three subcomponents is weighted in proportion to its importance in a given situation, with the constraint that the sum of the squared weights is unity.

The perception or observation of the actual performance of the ratee can be similarly partitioned into several aspects. The perception of one instance of job performance can be considered to be the sum of actual performance of the ratee in that instance, the bias of perception of the performance, and random error. The bias aspect can be further partitioned into three elements that are related to (a) the expected performance of the ratee, based on the true behavior component; (b) the residual effect (called areal bias) of all previous error and nonrelevant experiences that the rater has had with the ratee in situations identical or similar to the current one; and (c) an overall bias representing the residual effect of all possible areal bias effects. Each of these subcomponents of the perception of the actual performance is weighted in proportion to its importance as an effect on rating accuracy on a given occasion, with the constraint that the sum of the squared weights is unity.

The final stage of the rating process is the recall and reporting of previously perceived job performances. This recall is composed of the sum of the previous perceptions of performance, a systematic bias of recall, and random error of recall. The bias factor of recall may be considered to be composed of three elements analogous to those related to perception: true, areal bias, and overall bias. Again, each element of the recall component is weighted with regard to the magnitude of its effect on the rating, such that the sum of squared weights equals unity.

If it is assumed that the true performance level of a given ratee is constant, that a rater uses the same weights for the various components in all performance situations, that the areal bias of recall is equal to the areal bias of perception, and that the overall bias of recall is equal to the overall bias of perception, then the Wherry (Note 5) model of the rating process may be written in standard score units as

$$z_{X_R} = W_T z_T + W_{BRA} z_{BRA} + W_{BRO} z_{BRO} + W_I z_I + W_{EA} z_{EA} + W_{EP} z_{EP} + W_{ER} z_{ER}, \quad (1)$$

where z_{X_R} = recall of previous performance, z_T = true performance, z_{BRA} = areal bias of recall, z_{BRO} = overall bias of recall, z_I = environmental influences on performance, z_{EA} = random error of performance, z_{EP} = random error of perception, z_{ER} = random error of recall, and the Ws represent the weights assigned to the components, such that

$$W_T^2 + W_{BRA}^2 + W_{BRO}^2 + W_I^2 + W_{EA}^2 + W_{EP}^2 + W_{ER}^2 = 1.00.$$

From Equation 1 Wherry (Note 5) derived a total of 46 theorems and 24 corollaries concerning rating validity and reliability. Although we cannot describe in detail all of these theorems and corollaries, it is instructive to examine a few to better understand the substantive nature of the theory. For example, Wherry hypothesized that tasks in which the performance is maximally controlled by the ratee rather than by the work situation are more likely to result in accurate ratings be-

cause the z_I component will receive a relatively small weight. In addition, it was hypothesized that raters would be accurate in their ratings, if they had had much relevant contact with the ratee, consistent with later research by Freeberg (1969).

In other derivations from Equation 1, Wherry (Note 5) predicted that ratings gathered under research conditions would be more accurate than ratings gathered under administrative conditions and that increasing the number of observations, judgments, or raters will tend to reduce error. In particular, obtaining ratings about different areas of performance tends to reduce areal bias but has no effect on overall bias. Multiple raters may reduce both areal and overall bias, depending on the degree of correlation among irrelevant portions of rater–ratee contact.

An important theorem derived by Wherry (Note 5) stated that the reliability of a rating scale conveys little information about its validity, since the reliability of the measurement may be due to bias rather than to true performance. Since a frequent criterion of the value of a rating scale is its reliability, this theorem suggests that much rating research has been misdirected toward reliability rather than the important issues of validity and accuracy. More recently, several authors have stressed the necessity of using rating accuracy as the prime criterion in rating research (e.g., Borman, 1978; Vance et al., 1978).

The Wherry (Note 5) model also suggests several areas of research that have received only little attention to date. It is hypothesized that rating scales that refer to tasks that are operator paced will result in more accurate ratings than scales that refer to tasks that are environmentally controlled. Wherry also states that rater training should include instruction in observational techniques and in the keeping of written performance observations between rating periods. He also suggests that the requirement that ratings be seen by both the ratee and the rater's superior will tend to cancel out the effects of showing the ratings to only one or the other. Although many rating forms currently in use follow this suggestion, it has not been empirically tested.

From this brief exposure to the psychometric rating theory of Wherry (Note 5), it is evident that this approach has much to offer performance rating research. However, only a small number of studies concerned with performance ratings have cited Wherry (e.g., Barrett, 1966; Borman, 1978; Freeberg, 1969; Ross, 1966; Sharon & Bartlett, 1969; Centra, Note 9). This low citation level is probably due to the fact that the theory has never been published in the generally available literature. The theory should be carefully examined by performance rating researchers and tested empirically to a much greater extent than has already occurred.

The Wherry model has influenced the process model presented in Figure 2 principally in the consideration of the components of the cognitive element. The partitioning of the rating into variance components also has heuristic value. The identification of factors that influence ratings is simplified, when one considers that observation, storage, recall, and judgment constitute the rater's task rather than trying to deal with a unitary phenomenon of rating. Wherry also began a consideration of position characteristics, process variables, and so forth, that are more fully developed in the process model. In many ways, Figure 2 is a graphic representation of the variables discussed by Wherry. Figure 2 extends Wherry by considering the interactions among variables and the explicit cyclical nature of the total rating system. The process model developed above also suggests more clearly the many factors impinging on the cognitive element.

Unified Approach to Performance Rating

As indicated earlier, many researchers in the field of performance rating have recognized the need for models or frameworks to apply to the rating context. These models should integrate the various potential influences on performance descriptions into a single system. The process model that we propose in Figure 2 might be thought of as a superstructure of rating behavior. As such, it is primarily taxonomic in nature. Nevertheless, taxonomies represent a significant step in theory building. By identifying major com-

ponents of the system, hypothesis generation and testing is facilitated. Through the application of substantive models related to the process in question, higher level theories emerge. We feel that such is the case with the combination of our process model, implicit personality theory research, and Wherry's deductive propositions regarding variance components in ratings.

It is clear from even a cursory examination of the rating process that all information must ultimately pass through a cognitive filter represented by the rater. Multiple raters simply imply multiple filters that combine in some particular manner. Thus, one must understand the way in which environmental changes or constancies affect this cognitive operation called judgment. Organismic characteristics such as the sex, race, or age of the rater are peripheral. The more important questions relate to how cognitive operations are affected by group membership. The phenotypic characteristics of a rating format are theoretically less important than the interaction of these characteristics with cognitive operations that are involved in recording judgments concerning the behavior of others. Even data analytic procedures, such as factor analysis of multiple rating dimensions, presume that rater observations and the resulting ratings are veridical.

There are, of course, considerations beyond theory building. For the purposes of application and administration, it is useful to know the effect of contextual factors on ratings. Forewarned of certain systematic interactions between rater and ratee characteristics, practitioners are better able to develop equitable decision systems. The administrative implications of Wherry's model of rating variance are important. His propositions make it possible to determine what price is being paid for decreases in particular types of rating error (Wherry, Note 5). For example, the use of a rating scale with multiple items to be rated for each performance dimension increases the reliability of the performance rating for that dimension (due to random error reduction) but does not increase the relative proportion of true score to bias. Also, the addition of multiple raters, each evaluating the ratee on several performance dimensions, will reduce overall and areal bias in the composite rating only if the raters' irrelevant contacts with the ratee are at least somewhat different. If these contacts are the same, then there will be no reduction in bias. In addition to the administrative implications, Wherry clearly emphasized the importance of the cognitive operations of the rater, an emphasis only recently appearing in the applied literature on rating.

There is an enormous amount of work to be done, both inductive and deductive, in tying the propositions of person perception to the components of the process model. The propositions of Wherry regarding variance partitioning represent a procedure for accomplishing the initial steps of this integration.

Future Research Needs

The literature review suggests some global conclusions that can be drawn on the basis of available evidence. In one sense, rater and ratee characteristics are fixed—they cannot be changed as easily as can the format of a rating scale. Since little is known concerning the dynamics by which demographic characteristics, such as race and sex, affect ratings, it is difficult to imagine training programs that would eliminate specific biases peculiar to group membership. On the other hand, there is some indication that training in the use of a particular rating format is of value in reducing common rating errors. We must learn much more about the way in which potential raters observe, encode, store, retrieve, and record performance information, if we hope to increase the validity of ratings.

Given the superficiality of our knowledge of the cognitive processes of raters, we probably have gone as far as we can in improving rating formats. We know that the rater should have a clear understanding of the rating task, that the number of response categories should be limited, that the anchors on the scale should be rigorously developed, and that those anchors should be more than simple descriptive labels, such as poor, average, outstanding, and so forth. Nevertheless, even when all of these suggestions are taken into account, evidence suggests that their effect may be minimal. Data reviewed earlier indicate

that about 4%–8% of the variance in ratings can be explained on the basis of format. When one considers that the designs that yielded these estimates were seldom sensitive enough to identify true levels of performance, even this effect may be an overestimate. Thus, at least for the time being, we suggest a moratorium on format-related research.

Research in the area of statistical control of common rating errors has been sparse but encouraging. There are two distinct lines of research that might be fruitful. The first is the development of schemes for deriving residual performance scores—scores with rater–ratee context factors partialed out. Although this is a mechanical solution that implies no increase in the understanding of the rating process, it offers the possibility of simultaneously providing the practitioner with better numbers and the researcher with hypotheses. A second research line suggested by data analytic procedures is using ratings or judgments to derive cognitive maps or sets of raters. This would require some sophisticated designs directed toward understanding how individual raters construe their reality (Kelly, 1955).

Finally, our process model implies that the rater's experience with ratings affects the validity of those ratings. We know little or nothing about the effects that decisions based on current ratings have on future ratings. It is reasonable to assume that there are such effects. We can demonstrate that ratings for research purposes have different properties than ratings for administrative purposes. This implies that there is a feedback loop of some sort. Research in this area is long overdue. It is time to stop looking at the symptoms of bias in rating and begin examining potential causes.

Reference Notes

1. Wherry, R. J. *Control of bias in rating: Survey of the literature* (Tech. Rep. DA-49-0853 OSA 69). Washington, D.C.: Department of the Army, Personnel Research Section, September 1950.
2. Centra, J. A., & Linn, R. L. *Student points of view in ratings of college instruction* (ETS RB 73-60). Princeton, N.J.: Educational Testing Service, 1973.
3. Crooks, L. A. (Ed.). *An investigation of sources of bias in the prediction of job performance: A six-year study*. Princeton, N.J.: Educational Testing Service, 1972.
4. Centra, J. A. *Colleagues as raters of classroom instruction* (ETS RB 74-18). Princeton, N.J.: Educational Testing Service, 1974.
5. Wherry, R. J. *The control of bias in rating: A theory of rating* (Personnel Research Board Rep. 922). Washington, D.C.: Department of the Army, Personnel Research Section, February 1952.
6. Barnes, J. L. *Scaling assumptions in behaviorally anchored scale construction*. Paper presented at the meeting of the American Psychological Association, Toronto, August 1978.
7. Landy, F. J. Discussion. In S. Zedeck (Chair), *A closer look at performance appraisal through behavioral expectation scales*. Symposium presented at the meeting of the American Psychological Association, San Francisco, August 1977.
8. Landy, F. J., & Vance, R. J. *Statistical control of halo*. Unpublished manuscript, 1978. (Available from Frank J. Landy, Department of Psychology, Pennsylvania State University, University Park, Penn. 16802.)
9. Centra, J. A. *The influence of different directions on student ratings of instruction* (ETS RB 75-28). Princeton, N.J.: Educational Testing Service, 1975.

References

Amir, Y., Kovarsky, Y., & Sharan, S. Peer nominations as a predictor of multistage promotions in a ramified organization. *Journal of Applied Psychology*, 1970, *54*, 462–469.

Arvey, R. D., & Hoyle, J. C. A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. *Journal of Applied Psychology*, 1974, *59*, 61–68.

Baker, E. M., & Schuck, J. R. Theoretical note: Use of signal detection theory to clarify problems of evaluating performance in industry. *Organizational Behavior and Human Performance*, 1975, *13*, 307–317.

Barrett, R. S. Influence of supervisor's requirements on ratings. *Personnel Psychology*, 1966, *19*, 375–387. (a)

Barrett, R. S. *Performance rating*. Chicago: Science Research Associates, 1966. (b)

Barrett, R. S., Taylor, E. K., Parker, J. W., & Martens, L. Rating scale content: I. Scale information and supervisory ratings. *Personnel Psychology*, 1958, *11*, 333–346.

Bartlett, C. J. The relationship between self-ratings and peer ratings on a leadership behavior scale. *Personnel Psychology*, 1959, *12*, 237–246.

Bartol, K. M., & Butterfield, D. A. Sex effects in evaluating leaders. *Journal of Applied Psychology*, 1976, *61*, 446–454.

Bass, A. R., & Turner, J. N. Ethnic group differences in relationships among criteria of job performance. *Journal of Applied Psychology,* 1973, *57,* 101–109.

Bass, B. M. Reducing leniency in merit ratings. *Personnel Psychology,* 1956, *9,* 359–369.

Bass, B. M. Further evidence on the dynamic nature of criteria. *Personnel Psychology,* 1962, *15,* 93–97.

Bayroff, A. G., Haggerty, H. R., & Rundquist, E. A. Validity of ratings as related to rating techniques and conditions. *Personnel Psychology,* 1954, *7,* 93–114.

Bendig, A. W. A statistical report on a revision of the Miami instructor rating sheet. *Journal of Educational Psychology,* 1952, *43,* 423–429. (a)

Bendig, A. W. The use of student rating scales in the evaluation of instructors in introductory psychology. *Journal of Educational Psychology,* 1952, *43,* 167–175. (b)

Bendig, A. W. The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. *Journal of Applied Psychology,* 1953, *37,* 38–41.

Bendig, A. W. Reliability and number of rating scale categories. *Journal of Applied Psychology,* 1954, *38,* 38–40. (a)

Bendig, A. W. Reliability of short rating scales and the heterogeneity of the rated stimuli. *Journal of Applied Psychology,* 1954, *38,* 167–170. (b)

Berkshire, J. R., & Highland, R. W. Forced-choice performance rating—A methodological study. *Personnel Psychology,* 1953, *6,* 355–378.

Bernardin, H. J. Behavioral expectation scales versus summated scales: A fairer comparison. *Journal of Applied Psychology,* 1977, *62,* 422–427.

Bernardin, H. J. Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology,* 1978, *63,* 301–308.

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. A recomparison of behavioral expectation scales to summated scales. *Journal of Applied Psychology,* 1976, *61,* 564–570.

Bernardin, H. J., LaShells, M. B., Smith, P. C., & Alvares, K. M. Behavioral expectation scales: Effects of developmental procedures and formats. *Journal of Applied Psychology,* 1976, *61,* 75–79.

Bernardin, H. J., & Walter, C. S. Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology,* 1977, *62,* 64–69.

Berry, N. H., Nelson, P. D., & McNally, M. S. A note on supervisor ratings. *Personnel Psychology,* 1966, *19,* 423–426.

Bigoness, N. J. Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. *Journal of Applied Psychology,* 1976, *61,* 80–84.

Blanz, F., & Ghiselli, E. E. The mixed standard scale: A new rating system. *Personnel Psychology,* 1972, *25,* 185–199.

Blood, M. R. Spin-offs from behavioral expectation scale procedures. *Journal of Applied Psychology,* 1974, *59,* 513–515.

Blum, M. L., & Naylor, J. C. *Industrial psychology.* New York: Harper & Row, 1968.

Blumberg, H. H., DeSoto, C. B., & Kuethe, J. L. Evaluations of rating scale formats. *Personnel Psychology,* 1966, *19,* 243–259.

Booker, G. S., & Miller, R. W. A closer look at peer ratings. *Personnel,* 1966, *43(1),* 42–47.

Borman, W. C. The rating of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance,* 1974, *12,* 105–124.

Borman, W. C. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology,* 1975, *60,* 556–560.

Borman, W. C. Exploring upper limits of reliability and validity in performance ratings. *Journal of Applied Psychology,* 1978, *63,* 135–144.

Borman, W. C., & Dunnette, M. D. Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology,* 1975, *60,* 561–565.

Borman, W. C., & Vallon, W. R. A view of what can happen when behavioral expectation scales are developed in one setting and used in another. *Journal of Applied Psychology,* 1974, *59,* 197–201.

Borresen, H. A. The effects of instructions and item content on three types of ratings. *Educational and Psychological Measurement,* 1967, *27,* 855–862.

Brehmer, B. Cue utilization and cue consistency in multiple-cue probability learning. *Organizational Behavior and Human Performance,* 1972, *8,* 286–296.

Brown, E. M. Influence of training, method, and relationship on the halo effect. *Journal of Applied Psychology,* 1968, *52,* 195–199.

Brumback, G. A reply to Kavanagh. *Personnel Psychology,* 1972, *25,* 567–572.

Bruner, J. S., & Tagiuri, R. The perception of people. In G. Lindzey (Ed.), *Handbook of social psychology* (Vol. 2). Cambridge, Mass.: Addison-Wesley, 1954.

Buckner, D. N. The predictability of ratings as a function of interrater agreement. *Journal of Applied Psychology,* 1959, *43,* 60–64.

Burnaska, R. F., & Hollmann, T. D. An empirical comparison of the relative effects of rater response biases on three rating scale formats. *Journal of Applied Psychology,* 1974, *59,* 307–312.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81–105.

Campbell, D. T., Hunt, W. A., & Lewis, N. A. The relative susceptability of two rating scales to disturbances resulting from shifts in stimulus context. *Journal of Applied Psychology,* 1958, *42,* 213–217.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. The development and evaluation

of behaviorally based rating scales. *Journal of Applied Psychology,* 1973, *57,* 15–22.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., III, & Weick, K. E. *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill, 1970.

Carter, G. C. Measurement of supervisory ability. *Journal of Applied Psychology,* 1952, *36,* 393–395.

Cascio, W. F., & Valenzi, E. R. Behaviorally anchored rating scores: Effects of education and job experience of raters and ratees. *Journal of Applied Psychology,* 1977, *62,* 278–282.

Cotton, J., & Stoltz, R. E. The general applicability of a scale for rating research productivity. *Journal of Applied Psychology,* 1960, *44,* 276–277.

Cox, J. A., & Krumboltz, J. D. Racial bias in peer ratings of basic airmen. *Sociometry,* 1958, *21,* 292–299.

Cozan, L. W. Forced choice: Better than other rating methods? *Personnel Psychology,* 1959, *36,* 80–83.

Creswell, M. B. Effects of confidentiality on performance ratings of professional personnel. *Personnel Psychology,* 1963, *16,* 385–393.

Cronbach, L. J. Processes affecting scores on understanding of others and assuming "similarity." *Psychological Bulletin,* 1955, *52,* 177–193.

Dawes, R., & Corrigan, B. Linear models in decision making. *Psychological Bulletin,* 1974, *81,* 95–106.

DeCotiis, T. A. An analysis of the external validity and applied relevance of three rating formats. *Organizational Behavior and Human Performance,* 1977, *19,* 247–266.

DeJung, J. E., & Kaplan, H. Some differential effects of race of rater and combat attitude. *Journal of Applied Psychology,* 1962, *46,* 370–374.

Dickinson, T. L., & Tice, T. E. The discriminant validity of scales developed by retranslation. *Personnel Psychology,* 1977, *30,* 217–228.

Dunnette, M. D. A note on *the* criterion. *Journal of Applied Psychology,* 1963, *47,* 251–254.

Einhorn, H. J. Expert measurement and mechanical combination. *Organizational Behavior and Human Performance,* 1972, *7,* 86–106.

Elmore, P. B., & LaPointe, K. Effects of teacher sex and student sex on the evaluation of college instructors. *Journal of Educational Psychology,* 1974, *66,* 386–389.

Elmore, P. B., & LaPointe, K. A. Effect of teacher sex, student sex, and teacher warmth on the evaluation of college instructors. *Journal of Educational Psychology,* 1975, *67,* 368–374.

Farr, J. L., O'Leary, B. S., & Bartlett, C. J. Ethnic group membership as a moderator of the prediction of job performance. *Personnel Psychology,* 1971, *24,* 609–636.

Ferguson, L. W. The value of acquaintance ratings in criteria research. *Personnel Psychology,* 1949, *2,* 93–102.

Finn, R. H. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement,* 1972, *32,* 255–265.

Fiske, D. W., & Cox, J. A., Jr. The consistency of ratings by peers. *Journal of Applied Psychology,* 1960, *44,* 11–17.

Flanagan, J. C. The critical incident technique. *Psychological Bulletin,* 1954, *51,* 327–358.

Fox, H., & Lefkowitz, J. Differential validity: Ethnic group as a moderator in predicting job performance. *Personnel Psychology,* 1974, *27,* 209–223.

Frank, L. L., & Hackman, J. R. Effects of interviewer-interviewee similarity on interviewer objectivity in college admissions interviews. *Journal of Applied Psychology,* 1975, *60,* 356–360.

Freeberg, N. E. Relevance of rater-ratee acquaintance in the validity and reliability of ratings. *Journal of Applied Psychology,* 1969, *53,* 518–524.

Friedman, B. A., & Cornelius, E. T., III. Effect of rater participation in scale construction on the psychometric characteristics of two rating scale formats. *Journal of Applied Psychology,* 1976, *61,* 210–216.

Ghiselli, E. E., & Haire, M. The validation of selection tests in light of the dynamic character of criteria. *Personnel Psychology,* 1960, *13,* 225–231.

Goodale, J. G., & Burke, R. J. Behaviorally based rating scales need not be job specific. *Journal of Applied Psychology,* 1975, *60,* 389–391.

Gordon, L. V., & Medland, F. F. The cross-group stability of peer ratings of leadership potential. *Personnel Psychology,* 1965, *18,* 173–177.

Gordon, M. E. The effect of the correctness of the behavior observed on the accuracy of ratings. *Organizational Behavior and Human Performance,* 1970, *5,* 366–377.

Gordon, M. E. An examination of the relationship between the accuracy and favorability of ratings. *Journal of Applied Psychology,* 1972, *56,* 49–53.

Graham, W. K., & Calendo, J. T. Personality correlates of supervisory ratings. *Personnel Psychology,* 1969, *22,* 483–487.

Grant, D. L. A factor analysis of managers' ratings. *Journal of Applied Psychology,* 1955, *39,* 283–286.

Greenhaus, J. H., & Gavin, J. F. The relationship between expectancies and job behavior for white and black employees. *Personnel Psychology,* 1972, *25,* 449–455.

Grey, J., & Kipnis, D. Untangling the performance appraisal dilemma: The influence of perceived organizational context on evaluative processes. *Journal of Applied Psychology,* 1976, *61,* 329–335.

Guilford, J. P. *Psychometric methods* (2nd ed.). New York: McGraw-Hill, 1954.

Guilford, J. P., Christenson, R. R., Taaffe, G., & Wilson, R. C. Ratings should be scrutinized. *Educational and Psychological Measurement,* 1962, *22,* 439–447.

Guion, R. M. Criterion measurement and personnel judgments. *Personnel Psychology,* 1961, *14,* 141–149.

Guion, R. M. *Personnel testing.* New York: McGraw-Hill, 1965.

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness,

N. J. Race and sex as determinants of ratings by potential employers in a simulated work sampling task. *Journal of Applied Psychology,* 1974, *59,* 705–711.

Harari, O., & Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. *Journal of Applied Psychology,* 1973, *58,* 261–265.

Heneman, H. G., III. Comparisons of self- and superior ratings of managerial performance. *Journal of Applied Psychology,* 1974, *59,* 638–642.

Heron, A. The effects of real-life motivation on questionnaire response. *Journal of Applied Psychology,* 1956, *40,* 65–68.

Hollander, E. P. The reliability of peer nominations under various conditions of administration. *Journal of Applied Psychology,* 1957, *41,* 85–90.

Hollander, E. P. Validity of peer nominations in predicting a distant performance criterion. *Journal of Applied Psychology,* 1965, *49,* 434–438.

Huck, J. R., & Bray, D. W. Management assessment center evaluations and subsequent job performance of white and black females. *Personnel Psychology,* 1976, *29,* 13–30.

Isard, E. S. The relationship between item ambiguity and discriminating power in a forced-choice scale. *Journal of Applied Psychology,* 1956, *40,* 266–268.

Jacobson, M. B., & Effertz, J. Sex roles and leadership: Perceptions of the leaders and the led. *Organizational Behavior and Human Performance,* 1974, *12,* 383–396.

James, L. R. Criterion models and construct validity for criteria. *Psychological Bulletin,* 1973, *80,* 75–83.

Jay, R., & Copes, J. Seniority and criterion measures of job proficiency. *Journal of Applied Psychology,* 1957, *41,* 58–60.

Jenkins, G. D., & Taber, T. A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology,* 1977, *62,* 392–398.

Johnson, D. M. Reanalysis of experimental halo effects. *Journal of Applied Psychology,* 1963, *47,* 46–47.

Johnson, D. M., & Vidulich, R. N. Experimental manipulation of the halo effect. *Journal of Applied Psychology,* 1956, *40,* 130–131.

Jurgensen, C. E. Intercorrelations in merit rating traits. *Journal of Applied Psychology,* 1950, *34,* 240–243.

Jurgensen, C. E. Item weights in employee rating scales. *Journal of Applied Psychology,* 1955, *39,* 305–307.

Kane, J. S., & Lawler, E. E., III. Methods of peer assessment. *Psychological Bulletin,* 1978, *85,* 555–586.

Kaufman, G. G., & Johnson, J. C. Scaling peer ratings: An examination of the differential validities of positive and negative nominations. *Journal of Applied Psychology,* 1974, *59,* 302–306.

Kavanagh, M. J. The content issue in performance appraisal: A review. *Personnel Psychology,* 1971, *24,* 653–668.

Kavanagh, M. J. Rejoinder to Brumback "The content issue in performance appraisal: A review." *Personnel Psychology,* 1973, *26,* 163–166.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. Issues in managerial performance: Multitrait-multimethod analysis of ratings. *Psychological Bulletin,* 1971, *75,* 34–49.

Kay, B. R. The use of critical incidents in a forced-choice scale. *Journal of Applied Psychology,* 1959, *43,* 269–270.

Keaveny, T. J., & McGann, A. F. A comparison of behavioral expectation scales and graphic rating scales. *Journal of Applied Psychology,* 1975, *60,* 695–703.

Kelly, G. A. *The psychology of personal constructs.* New York: Norton, 1955.

Kirchner, W. K. Relationships between supervisory and subordinate ratings for technical personnel. *Journal of Industrial Psychology,* 1965, *3,* 57–60.

Kirchner, W. K., & Reisberg, D. J. Differences between better and less effective supervisors in appraisal of subordinates. *Personnel Psychology,* 1962, *15,* 295–302.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. *Testing and fair employment.* New York: New York University Press, 1968.

Klieger, W. A., & Mosel, J. N. The effect of opportunity to observe and rater status on the reliability of performance ratings. *Personnel Psychology,* 1953, *6,* 57–64.

Klimoski, R. J., & London, M. Role of the rater in performance appraisal. *Journal of Applied Psychology,* 1974, *59,* 445–451.

Klores, M. S. Rater bias in forced-distribution ratings. *Personnel Psychology,* 1966, *19,* 411–421.

Koltuv, B. Some characteristics of intrajudge trait intercorrelations. *Psychological Monographs,* 1962, *76*(33, Whole No. 552).

Kraut, A. J. Prediction of managerial success by peer and training-staff ratings. *Journal of Applied Psychology,* 1975, *60,* 14–19.

Landy, F. J., Barnes, J., & Murphy, K. Correlates of perceived fairness and accuracy in performance appraisal. *Journal of Applied Psychology,* 1978, *63,* 751–754.

Landy, F. J., & Farr, J. L. Police performance appraisal. JSAS *Catalog of Selected Documents in Psychology,* 1976, *6,* 83. (Ms. No. 1315)

Landy, F. J., Farr, J. L., Saal, F. G., & Freytag, W. R. Behaviorally anchored scales for rating the performance of police officers. *Journal of Applied Psychology,* 1976, *61,* 752–758.

Landy, F. J., & Guion, R. M. Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance,* 1970, *5,* 93–103.

Landy, F. J., & Trumbo, D. A. *The psychology of work behavior* (Rev. ed.). Homewood, Ill.: Dorsey Press, in press.

Latham, G. P., & Wexley, K. N. Behavioral observation scales for performance appraisal purposes. *Personnel Psychology,* 1977, *30,* 255–268.

Latham, G. P., Wexley, K. N., & Pursell, E. D. Training managers to minimize rating errors in the

observation of behavior. *Journal of Applied Psychology,* 1975, *60,* 550–555.

Lawler, E. E., III. The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology,* 1967, *51,* 369–381.

Lay, C. H., Burron, B. F., & Jackson, D. N. Base rates and informational value in impression formation. *Journal of Personality and Social Psychology,* 1973, *28,* 390–395.

Lee, D., & Alvares, K. Effect of sex on descriptions and evaluations of supervisory behavior in a simulated industrial setting. *Journal of Applied Psychology,* 1977, *62,* 405–410.

Lepkowski, J. R. Development of a forced-choice rating scale for engineer evaluation. *Journal of Applied Psychology,* 1963, *47,* 87–88.

Leventhal, L., Perry, R. P., & Abrami, P. C. Effect of lecturer quality and student perception of lecturer's experience on teacher ratings. *Journal of Educational Psychology,* 1977, *69,* 360–374.

Lewin, A. Y., & Zwany, A. Peer nominations: A model, literature critique and a paradigm for research. *Personnel Psychology,* 1976, *29,* 423–447.

Lewis, N. A., & Taylor, J. A. Anxiety and extreme response preferences. *Educational and Psychological Measurement,* 1955, *15,* 111–116.

Lissitz, R. W., & Green, S. B. Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology,* 1975, *60,* 10–13.

London, M., & Poplawski, J. R. Effects of information on stereotype development in performance appraisal and interview context. *Journal of Applied Psychology,* 1976, *61,* 199–205.

Lopez, F. M. *Evaluating employee performance.* Chicago: Public Personnel Association, 1968.

Lovell, G. D., & Haner, C. F. Forced-choice applied to college faculty rating. *Educational and Psychological Measurement,* 1955, *15,* 291–304.

Maas, J. B. Patterned scaled expectation interview: Reliability studies on a new technique. *Journal of Applied Psychology,* 1965, *59,* 431–433.

Madden, J. M., & Bourdon, R. D. Effects of variations in rating scale format on judgment. *Journal of Applied Psychology,* 1964, *48,* 147–151.

Mandell, M. M. Supervisory characteristics and ratings: A summary of recent research. *Personnel Psychology,* 1956, *32,* 435–440.

Meyer, H. H. Methods for scoring a check-list type rating scale. *Journal of Applied Psychology,* 1951, *35,* 46–49.

Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review,* 1956, *63,* 81–97.

Miner, J. B. Management appraisal: A review of procedures and practices. In H. L. Tosi, R. J. House, & M. D. Dunnette (Eds.), *Managerial motivation and compensation.* East Lansing, Mich.: Michigan State University, Graduate School of Business Administration, 1972.

Mischel, H. N. Sex bias in the evaluation of professional achievements. *Journal of Educational Psychology,* 1974, *66,* 157–166.

Moore, L. F., & Lee, A. J. Comparability of interviewer, group, and individual interview ratings. *Journal of Applied Psychology,* 1974, *59,* 163–167.

Mullins, C. J., & Force, R. C. Rater accuracy as a generalized ability. *Journal of Applied Psychology,* 1962, *46,* 191–193.

Myers, J. H. Removing halo from job evaluation factor structure. *Journal of Applied Psychology,* 1965, *49,* 217–221.

Naylor, J. C., & Wherry, R. J., Sr. The use of simulated stimuli and the "JAN" technique to capture and cluster the policies of raters. *Educational and Psychological Measurement,* 1965, *25,* 969–986.

Norman, W. T., & Goldberg, L. R. Rater, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology,* 1966, *4,* 681–691.

Obradovic, J. Modification of the forced-choice method as a criterion of job proficiency. *Journal of Applied Psychology,* 1970, *54,* 228–233.

Overall, J. E. Reliability of composite ratings. *Educational and Psychological Measurement,* 1965, *25,* 1011–1022.

Parker, J. W., Taylor, E. K., Barrett, R. S., & Martens, L. Rating scale content: 3. Relationship between supervisory and self-ratings. *Personnel Psychology,* 1959, *12,* 49–63.

Passini, F. T., & Norman, W. T. Ratee relevance in peer nominations. *Journal of Applied Psychology,* 1969, *53,* 185–187.

Paterson, D. G. The Scott Company graphic rating scale. *Journal of Personnel Research,* 1922, *1,* 361–376.

Peters, D. L., & McCormick, E. J. Comparative reliability of numerically anchored versus job-task anchored rating scales. *Journal of Applied Psychology,* 1966, *50,* 92–96.

Prien, E. P. Dynamic character of criteria: Organizational change. *Journal of Applied Psychology,* 1966, *50,* 501–504.

Ritti, R. R. Control of "halo" in factor analysis of a supervisory behavior inventory. *Personnel Psychology,* 1964, *17,* 305–318.

Ronan, W. W., & Prien, E. P. *Toward a criterion theory: A review and analysis of research and opinion.* Greensboro, N.C.: Creativity Institute of the Richardson Foundation, 1966.

Rosen, B., & Jerdee, T. H. The influence of sex role stereotypes on evaluations of male and female supervisory behavior. *Journal of Applied Psychology,* 1973, *57,* 44–48.

Ross, P. F. Reference groups in man-to-man job performance rating. *Personnel Psychology,* 1966, *19,* 115–142.

Rothaus, P., Morton, R. B., & Hanson, P. G. Performance appraial and psychological distance. *Journal of Applied Psychology,* 1965, *49,* 48–54.

Rothe, H. F. The relation of merit ratings to length of service. *Personnel Psychology,* 1949, *2,* 237–242.

Rotter, G. S., & Tinkleman, V. Anchor effects in the development of behavior rating scales. *Educational and Psychological Measurement,* 1970, *30,* 311–318.

Ryan, F. J. Trait ratings of high school students by

teachers. *Journal of Educational Psychology,* 1958, *49,* 124–128.

Saal, F. E., & Landy, F. J. The mixed standard rating scale: An evaluation. *Organizational Behavior and Human Performance,* 1977, *18,* 19–35.

Schein, V. E. The relationship between sex role stereotypes and requisite management characteristics. *Journal of Applied Psychology,* 1973, *57,* 95–100.

Schein, V. E. Relationships between sex role stereotypes and requisite management characteristics among female managers. *Journal of Applied Psychology,* 1975, *60,* 340–344.

Schmidt, F. L., & Johnson, R. H. Effect of race on peer ratings in an industrial setting. *Journal of Applied Psychology,* 1973, *57,* 237–241.

Schmidt, F. L., & Kaplan, L. B. Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology,* 1971, *24,* 419–434.

Schmitt, N., & Hill, T. Sex and race composition of assessment center groups as a determinant of peer and assessor ratings. *Journal of Applied Psychology,* 1977, *62,* 261–264.

Schneider, D. E., & Bayroff, A. G. The relationship between rater characteristics and validity of ratings. *Journal of Applied Psychology,* 1953, *37,* 278–280.

Schneider, D. J. Implicit personality theory. *Psychological Bulletin,* 1973, *79,* 294–309.

Schneier, C. E. Operational utility and psychometric characteristics of behavioral expectation scales: A cognitive reinterpretation. *Journal of Applied Psychology,* 1977, *62,* 541–548.

Schultz, D. G., & Siegel, A. I. The analysis of job performance by multi-dimensional scaling techniques. *Journal of Applied Psychology,* 1964, *48,* 329–335.

Schwab, D. P., Heneman, H. G., III, & DeCotiis, T. Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology,* 1975, *28,* 549–562.

Schwartz, D. J. A job sampling approach to merit system examining. *Personnel Psychology,* 1977, *30,* 175–185.

Scott, W. E., Jr., & Hamner, W. C. The influence of variations in performance profiles on the performance evaluation process: An examination of the validity of the criterion. *Organizational Behavior and Human Performance,* 1975, *14,* 360–370.

Sharon, A. T. Eliminating bias from student ratings of college instructors. *Journal of Applied Psychology,* 1970, *54,* 278–281.

Sharon, A. T., & Bartlett, C. J. Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology,* 1969, *22,* 251–263.

Smith, P. C. Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand McNally, 1976.

Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology,* 1963, *47,* 149–155.

Springer, D. Ratings of candidates for promotion by co-workers and supervisors. *Journal of Applied Psychology,* 1953, *37,* 347–351.

Staugas, L., & McQuitty, L. L. A new application of forced-choice ratings. *Personnel Psychology,* 1950, *3,* 413–424.

Stone, E., Rabinowitz, S., & Spool, M. D. Effect of anonymity on student evaluations of faculty performance. *Journal of Educational Psychology,* 1977, *69,* 274–280.

Suci, G. J., Vallance, T. R., & Glickman, A. S. A study of the effects of "likingness" and level of objectivity on peer rating reliabilities. *Educational and Psychological Measurement,* 1956, *16,* 147–152.

Svetlik, B., Prien, E., & Barrett, G. Relationships between job difficulty, employee's attitude toward his job, and supervisory ratings of the employee effectiveness. *Journal of Applied Psychology,* 1964, *48,* 320–324.

Tajfel, H., & Wilkes, A. L. Salience of attributes and commitment to extreme judgments in perception of people. *British Journal of Social and Clinical Psychology,* 1963, *2,* 40–49.

Taylor, E. K., & Hastman, R. Relation of format and administration to the characteristics of graphic scales. *Personnel Psychology,* 1956, *9,* 181–206.

Taylor, E. K., Parker, J. W., Martens, L., & Ford, G. L. Supervisory climate and performance ratings, an exploratory study. *Personnel Psychology,* 1959, *12,* 453–468.

Taylor, E. K., Schneider, D. E., & Clay, H. C. Short forced-choice ratings work. *Personnel Psychology,* 1954, *7,* 245–252.

Taylor, E. K., & Wherry, R. J. A study of leniency in two rating systems. *Personnel Psychology,* 1951, *4,* 39–47.

Taylor, J. B., Haeffele, E., Thompson, P., & O'Donoghue, C. Rating scales as measures of clinical judgment: 2. The reliability of example-anchored scales under conditions of rater heterogeneity and divergent behavior sampling. *Educational and Psychological Measurement,* 1970, *30,* 301–310.

Terborg, J. R., & Ilgen, D. R. A theoretical approach to sex discrimination in traditionally masculine occupations. *Organizational Behavior and Human Performance,* 1975, *13,* 352–376.

Toole, D. L., Gavin, J. F., Murdy, L. B., & Sells, S. B. The differential validity of personality, personal history, and aptitude data for minority and nonminority employees. *Personnel Psychology,* 1972, *25,* 661–672.

Vance, R. J., Kuhnert, K. W., & Farr, J. L. Interview judgments: Using external criteria to compare behavioral and graphic scale ratings. *Organizational Behavior and Human Performance,* 1978, *22,* 279–294.

Wagner, E. E., & Hoover, T. O. The influence of technical knowledge on position error in ranking. *Journal of Applied Psychology,* 1974, *59,* 406–407.

Waters, L. K., & Waters, C. W. Peer nominations as predictors of short-term sales performance. *Journal of Applied Psychology,* 1970, *54,* 42–44.

Wells, W. D., & Smith, G. Four semantic rating scales compared. *Journal of Applied Psychology,* 1960, *44,* 393–397.

Wexley, K. N., Sanders, R. E., & Yukl, G. A. Training interviewers to eliminate contrast effects in employment interviews. *Journal of Applied Psychology,* 1973, *57,* 233–236.

Wherry, R. J., & Naylor, J. C. Comparison of two approaches—JAN and PROF—for capturing rater strategies. *Educational and Psychological Measurement,* 1966, *26,* 267–286.

Whitla, D. K., & Tirrell, J. E. The validity of ratings of several levels of supervisors. *Personnel Psychology,* 1953, *6,* 461–466.

Willingham, W. W. Interdependence of successive absolute judgments. *Journal of Applied Psychology,* 1958, *42,* 416–418.

Windle, C. D., & Dingman, H. F. Interrater agreement and predictive validity. *Journal of Applied Psychology,* 1960, *44,* 203–204.

Wright, P. The harassed decision maker: Time pressures, distractions, and the use of evidence. *Journal of Applied Psychology,* 1974, *59,* 555–561.

Zavala, A. Development of the forced-choice rating scale technique. *Psychological Bulletin,* 1965, *63,* 117–124.

Zedeck, S., & Baker, H. T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance,* 1972, *7,* 457–466.

Zedeck, S., Imparato, N., Krausz, M., & Oleno, T. Development of behaviorally anchored rating scales as a function of organizational level. *Journal of Applied Psychology,* 1974, *59,* 249–252.

Zedeck, S., Jacobs, R., & Kafry, D. Behavioral expectations: Development of parallel forms and analysis of scale assumptions. *Journal of Applied Psychology,* 1976, *61,* 112–115.

Zedeck, S., & Kafry, D. Capturing rater policies for processing evaluation data. *Organizational Behavior and Human Performance,* 1977, *18,* 269–294.

Zedeck, S., Kafry, D., & Jacobs, R. Format and scoring variations in behavioral expectation evaluations. *Organizational Behavior and Human Performance,* 1976, *17,* 171–184.

Received August 23, 1978 ■