Psychological Bulletin
1980, Vol. 88, No. 2, 413-428

Rating the Ratings: Assessing the Psychometric Quality of Rating Data

Frank E. Saal, Ronald G. Downey, and Mary Anne Lahey
Kansas State University

Research concerned with evaluating the psychometric qualities of data in the form of ratings (rating errors) has been plagued with conceptual and operational confusion and inconsistency. Following a brief historical survey of such research, inconsistencies in definitions, quantifications, and methodologies are documented in a review of more than 20 relevant articles published in *Journal of Applied Psychology, Organizational Behavior and Human Performance,* and *Personnel Psychology* from 1975 through 1977. Empirical implications of these inconsistencies are discussed, and a revised typology of rating criteria, combined with a multivariate analytic approach, is suggested.

The use of a pencil-and-paper instrument by one individual to communicate judgments about one or more aspects of other individuals is common in contemporary society. Few social sciences, if any, make greater use of such judgmental information than psychology. Although many of these judgments are in the form of rankings, paired comparisons, or checklists, the most popular judgmental measure is the rating scale (Guion, 1965; Landy & Trumbo, 1976). Industrial psychologists rely on ratings in the contexts of job performance appraisal, personnel selection (especially interviews and assessment centers), and leadership evaluation, to name but a few. Educational and school psychologists often use teachers' ratings of their students' behavior in the classroom (Frandsen, 1967). Clinical psychologists point to ratings as a useful behavioral assessment technique in their efforts to better understand and serve their patients or clients (Phares, 1979). The rating scale clearly pervades the various domains of applied psychology and is widely used throughout our society as well.

Equally pervasive, however, are the suspicions and criticisms associated with the use of rating scales and the information they provide. Most of the reservations, regardless of how elegantly phrased, reflect fears that rating scale data are subjective (emphasizing, of course, the undesirable connotations of subjectivity), biased, and at worst, purposefully distorted. In the face of these often warranted misgivings, it is not surprising that psychologists and other professionals who rely on rating data have devoted much time and effort to the development of techniques and procedures for assessing the veracity and general psychometric qualities of their judgmental measures. Although most of the strategies that emerged were designed to identify *rating errors*—inadequacies of one sort or another in the ratings—not all of the suggested criteria for gauging the quality of ratings have been couched in negative terms.

In the attempt to alter the perception of ratings as "too subjective" by developing indices or criteria of rating quality, the research has unfortunately added to the confusion. First, there is less than unanimous agreement regarding conceptual definitions for

---

The first two authors contributed equally to the writing of this article.

We are grateful to Dallas Johnson for his statistical and computer programming assistance, to Tony Dubitsky for his contributions to our survey of the research literature, and to Robert McIntyre for his very helpful comments on an earlier draft of this article.

Requests for reprints should be sent to Frank E. Saal or to Ronald G. Downey, Department of Psychology, Kansas State University, Manhattan, Kansas 66506.

 0033-2909/80/8802-0413$00.75

several of the criteria of rating quality. Second, there is even less agreement regarding the operational definitions (statistical indices) for those criteria (Downey & Saal, Note 1). Third, different researchers have used different research designs or data collection procedures with inherently limited capabilities for aggregating data and yielding particular statistical indices of rating quality. It is therefore easy to find two or more studies in the literature that use the same label for a particular criterion of rating quality (e.g., *halo*) even though the conceptual and operational definitions of that particular rating error are not identical and the data collection strategies are sufficiently different to preclude calculation of similar statistical indices.

The present article has four objectives: (a) to briefly describe some of the early literature designed to identify and develop indices of criteria for gauging the quality of rating scale data, (b) to document the inconsistencies and potential confusion that characterize both conceptual and operational definitions of particular rating errors and other quality criteria, (c) to suggest the implications of these discrepant definitions by referring to empirical comparisons of rating data obtained using two different rating scale formats, and (d) to describe and offer for consideration an integrated system of definitions and data collection procedures for assessing the psychometric properties of rating scale data.

Early Literature

Although references to the phenomenon appeared in the work of Wells (1909) and Webb (1915), one of the first landmarks in the area of rating criteria was Thorndike's (1920) research on halo—a rater's tendency to think of a person as being generally good or generally inferior—which would yield relatively high intercorrelations between ostensibly different dimensions of performance or behavior. Immediately following Thorndike's contribution appeared the prodigious work of Rugg (1921a, 1921b, 1922a, 1922b), whose investigations of Army personnel ratings closely paralleled work in the area of group intelligence testing (Army Alpha and Beta tests). Using a form of interrater agreement as his primary evaluation criterion, Rugg (1922a) concluded that more accurate ratings can be obtained by averaging several judgments and that the inadequacy of a single rater's estimate was due to "general attitudes toward our associates and subordinates" (p. 39)—halo.

Kingsbury (1922) focused on three of the major criteria used today for evaluating the quality of ratings: (a) halo, as discussed by Thorndike (1920); (b) high–low raters, the forerunner of *leniency–severity*; and (c) fear of making distinctions, the precursor of *central tendency* or *range restriction*. He made no attempt, however, to quantify these rating errors.

The term *lenient* seems to have been coined by Kneeland (1929) to describe the tendency of raters to "rate well above the midpoint of the scales used" (p. 356), as indicated by average ratings over all ratees. Ford (1931) adopted Kneeland's terminology and contributed the term *severe* to describe raters who concentrate on the low end of the scale. Ford also suggested a strategy for eliminating the effects of these two tendencies from obtained ratings.

Newcomb (1931) described a different kind of rating error, *logical error*, wherein strong relationships between intraindividual behaviors were "presumed to spring from logical presuppositions in the minds of the raters" (p. 288). The logical error bears a striking resemblance to the contemporary idea of implicit personality theory (Schneider, 1973). Murray (1938) identified the *contrast error*, which was described as a tendency of raters to compare the ratees with themselves (the raters). Finally, Stockford and Bissell (1949) reported that the degree of correlation between traits seemed to vary as a function of the physical distance between those traits on the rating form itself and dubbed this *proximity error*.

Of course, the most systematic attempt to explore, define, and quantify various rating errors was Guilford's (1954) treatment of rating scales in his classic *Psychometric Methods*.

These few brief paragraphs certainly do not capture the richness of the literature devoted to rating criteria during the past three quarters of a century.[1] They do, however, sketch a historical context in which to appreciate the state of the art as it currently exists. The remainder of this article is devoted to a detailed description of that state, as well as to specific suggestions for alleviating some of the confusion that currently characterizes it.

Conceptual and Operational Definitions of Rating Criteria

The variety of conceptual and operational definitions of rating criteria that can be found in recent research literature is documented in this section. Specifically, these definitions emerged from a review of more than 20 articles that dealt either directly or indirectly with rating errors or with other indices of rating quality, published in the *Journal of Applied Psychology, Organizational Behavior and Human Performance,* or *Personnel Psychology* from 1975 through 1977. Those articles, along with the criteria of rating quality examined in each (and the data collection procedures used), are listed in Table 1.

Halo

Conceptual definitions of halo are relatively free of the inconsistency that seems to characterize the abstract definitions of some of the other rating criteria. Alternatively defined as (a) a tendency to attend to a global impression of each ratee rather than to carefully distinguish among levels of different performance dimensions (Borman, 1975), (b) a rater's inability or unwillingness to distinguish among the dimensions of a given ratee's job behavior (DeCotiis, 1977), or (c) a tendency to place a given ratee at the same level on different dimensions (Bernardin, 1977), the halo effect is consistently conceptualized as a rater's failure to discriminate among conceptually distinct and potentially independent aspects of a ratee's behavior.

Nisbett and Wilson (1977), however, whose research focused on the dynamics of the halo effect, sounded a warning to those who might be lulled into a state of complacency by this apparent conceptual consistency. Central to their argument is the distinction between a strong interpretation of halo and a weaker interpretation. Briefly stated, a strong interpretation implies that global evaluations actually alter more specific evaluations of particular performance dimensions for which the rater has information fully sufficient to permit independent assessment. A weaker interpretation implies merely that a global evaluation "might color presumptions about specific traits or influence interpretation of the meaning or affective value of *ambiguous* trait information" (p. 250; italics added for emphasis). The authors observed that the strong interpretation of the halo effect, which is consistent with Thorndike's (1920) earliest theorizing about the phenomenon, can only be demonstrated through an experimental approach (see, e.g., Johnson, 1963; Johnson & Vidulich, 1956).

The literature review summarized in Table 1 revealed four different operational definitions of the halo effect. One approach examines the intercorrelations among different dimension ratings, using ratee scores (over raters) for each dimension as data points (Keaveny & McGann, 1975). Higher correlations suggest less discrimination among different aspects of behavior and thus more halo. A second approach, an extension of the first, focuses on the results of a factor or principal-component analysis of the dimension intercorrelation matrix (Kraut, 1975; see, also, Blanz & Ghiselli, 1972): the fewer factors or principal components that emerge, the greater the halo. Emergence of a single factor or principal component that explains a sizable proportion of the rating variance consti-

[1] Although concern with rating errors apparently awaited the turn of the century, the use of psychological rating scales did not. Although Galton is commonly credited with introducing rating scales as psychological measuring devices (Garrett & Schneck, 1933), an interesting historical note by Ellson and Ellson (1953) suggested that Robert Owen (who founded the New Harmony Colony in 1825) was using "highly developed" psychological rating scales to measure children's capabilities "when Galton was a child."

Table 1
Studies on the Quality of Rating Data

Study	Quality criteria				Type of data matrix
	H	L/S	CT/RR	IR/A	
Bernardin (1977)	x	x	y	x	Rater × Dimension (15)
Bernardin, Alvares, & Cranny (1976)	x	x	y	x	Rater × Dimension (27)
Bernardin, LaShells, Smith, & Alvares (1976)	x	x	y	x	Rater × Dimension (20)
Bernardin & Walter (1977)	x	x	y	x	Rater × Dimension (13)
Borman (1975)	x			x	Rater × Dimension (6)
Borman (1977)	x	x	x		Rater × Ratee × Dimension
Borman & Dunnette (1975)	x	x	y	x	Rater × Ratee × Dimension (partial)
Cascio & Valenzi (1977)	x	x			Rater × Ratee × Dimension (partial)
DeCotiis (1977)	x	x	x	x	Rater × Ratee × Dimension
Dickinson & Tice (1977)	x				Rater × Ratee × Dimension (partial)
Finley, Osburn, Dubin, & Jeanneret (1977)	x	x	y	x	Rater × Ratee × Dimension (partial)
Friedman & Cornelius (1976)	x	x		x	Rater × Ratee × Dimension
Heneman, Schwab, Huett, & Ford (1975)	x	x	y	x	Rater × Ratee × Dimension
Keaveny & McGann (1975)	x	x		x	Rater × Dimension (4)
Kraut (1975)	x				Rater × Ratee × Dimension
Landy, Farr, Saal, & Freytag (1976)	x	x	x	x	Rater × Ratee × Dimension (partial)
Latham, Wexley, & Pursell (1975)	x				Rater × Ratee × Dimension
Motowidlo & Borman (1977)	x	x	x	x	Rater × Dimension (47)
Rizzo & Frank (1977)	x				Rater × Ratee × Dimension
Saal & Landy (1977)	x	x		x	Rater × Ratee × Dimension (partial)
Schneier (1977)	x	x	x		Rater × Ratee × Dimension (partial)
Zedeck, Kafry, & Jacobs (1976)	x	x		x	Rater × Dimension

Note. H = halo; L/S = leniency or severity; CT/RR = central tendency or range restriction; IR/A = interrater reliability or agreement. An x indicates that a study used a particular criterion of rating quality. A y indicates that although the terms *central tendency* or *range restriction* were not used, these studies used a conceptually similar or identical criterion of rating quality and described it as *discriminability, discrimination across ratees, differentiation among ratees,* or *variability of ratings.* The number of separate Rater × Dimension data matrices examined by the author(s) is given in parentheses. With the exceptions of Landy, Farr, Saal, and Freytag (1976) and Motowidlo and Borman (1977), all of these studies compared two or more rating scale formats (e.g., behaviorally anchored vs. graphic rating scales) or two or more levels of some other treatment variable (e.g., rater training).

tutes the limiting case and reflects a maximal halo effect.

A third approach concentrates on the variance (or standard deviation) associated with a particular rater's ratings of a particular ratee across all the performance dimensions (Bernardin & Walter, 1977). Less dispersion among the dimension ratings, evidenced by smaller standard deviation or variance estimates, indicates a greater halo effect. A fourth approach is based on a Rater × Ratee × Dimension analysis of variance (ANOVA) (Dickinson & Tice, 1977; see, also, Guilford, 1954; Stanley, 1961; Willingham & Jones, 1958). Emergence of a statistically significant Rater × Ratee interaction, especially one that explains a sizable proportion of the rating variance, has been interpreted in the literature

as a sign of halo, although Stanley and Willingham and Jones indicated that this is something of an oversimplification.

Leniency or Severity

Conceptual definitions of leniency or severity in the articles reviewed included the following: (a) a tendency to assign a higher or lower rating to an individual than is warranted by that ratee's behavior (Saal & Landy, 1977); (b) a response set attributed to "easy" or "hard-nosed" raters whose ratings are consistently higher or lower than is warranted, given some external criterion of known true performance level (DeCotiis, 1977); (c) a shift in mean ratings from the midpoint of the rating scale in the favorable or unfavorable direction (Bernardin, LaShells, Smith, & Alvares, 1976; see, also, Sharon & Bartlett, 1969); and (d) a rating level effect (Borman, 1977). Although the notion that ratings are consistently too high or too low pervades each of these four conceptualizations, their underlying assumptions are noticeably different. The first two refer explicitly or implicitly to a true performance score—an "eye-of-God-reality" (Lumsden, 1976) to which mortal psychologists must remain forever blind. The third suggests that average performance levels should coincide with rating scale midpoints, an assumption employers strive to invalidate with effective personnel selection, placement, and training programs. The fourth description, although free of questionable assumptions, may appear to be sterile and uninformative.

The articles listed in Table 1 offered three different operational definitions of leniency or severity. By far the most popular approach entails simple comparisons of average dimension ratings with the midpoint(s) of the rating scale(s) used (Bernardin, Alvares, & Cranny, 1976). Mean dimension ratings that exceed the midpoint are thought to reflect leniency, whereas mean ratings below the midpoint are thought to reflect severity. A second and much less popular approach is based on a Rater × Ratee × Dimension ANOVA in which a statistically significant rater main effect, especially one that explains a sizable proportion of the rating variance, is interpreted as leniency or severity (Friedman & Cornelius, 1976). A third approach, used even more infrequently, examines the degree of skewness that characterizes frequency distributions of dimension ratings (Landy, Farr, Saal, & Freytag, 1976). Significant negative skewness is thought to reflect leniency, whereas significant positive skewness is interpreted as severity.

Central Tendency and Restriction of Range

A third rating error, often discussed along with halo and leniency or severity in industrial psychology textbooks (Landy & Trumbo, 1976; McCormick & Tiffin, 1974; Wexley & Yukl, 1977; Zedeck & Blood, 1974), is central tendency. Although DeCotiis (1977) glibly defined central tendency as a rater's unwillingness to go out on the proverbial limb in either the favorable or unfavorable direction, definitions of this particular criterion of rating quality are limited almost exclusively to textbook discussions. The term *central tendency* appeared only rarely in the literature reviewed (Table 1). The textbook conceptualizations are commonly more formal statements of DeCotiis's description, emphasizing raters' reluctance to make extreme judgments about other individuals.

Rather than discussing central tendency, several of the articles listed in Table 1 examined *restriction of range*, a similar criterion of rating quality, defined as the extent to which obtained ratings discriminate among different ratees in terms of their respective performance levels (Motowidlo & Borman, 1977; Schneier, 1977). Since performance ratings that fail to adequately discriminate among ratees surely do not expedite administrative decision making and have "a potentially serious effect on attempts to establish empirical validity for selection devices" (Landy & Trumbo, 1976, p. 113), this is certainly a legitimate criterion of the quality of ratings.

All would be well, were it not so tempting to indiscriminately use the terms *central tendency* and *restriction of range* (or discriminability) as synonyms. This is not to

say that they cannot be so used; it depends on one's definition of central tendency.

Korman (1971), whose description is a typical one, defined central tendency as "the tendency to rate all rating objects around the 'middle' or mean of a rating continuum and not to use the extremes" (pp. 180–181). If the mean or average rating coincides with the middle or midpoint of the rating continuum (scale), there is no confusion. In this case raters are apparently unwilling to make extremely positive or negative judgments about ratees, and this unwillingness results in a restriction in the range of the obtained ratings in the vicinity of the scale midpoint. Unfortunately, the average rating often does not coincide with the scale midpoint. (See the earlier discussion of leniency and severity.) In such cases, in which the mean of the ratings may be well above or below the middle of the scale, the ratings could be described as being relatively free of central tendency, at least according to one interpretation of Korman's definition. Restriction of range, however, may persist. Range restriction around a lenient or severe mean rating is no less detrimental to administrative decision making and establishing the empirical validity of selection devices than is range restriction around a mean rating that happens to coincide with the midpoint of the rating scale.

It is therefore advisable to clearly distinguish between central tendency and range restriction. Central tendency should be used exclusively to describe situations in which ratings are clustered about the midpoint of the rating scale, reflecting raters' reluctance to use either of the extreme ends of the continuum (following Korman's, 1971, definition). Restriction of range can be used in situations in which ratings are clustered about any point on the rating continuum, be it a favorable (lenient) point, an unfavorable (severe) point, or the midpoint (central tendency) on the rating scale. Thus, central tendency implies range restriction; but the converse is not true, since range restriction may reflect leniency, severity, or central tendency.

The literature review (Table 1) revealed several different operational definitions of central tendency and range restriction (or lack of discriminability). The most common approach involves nothing more than calculating the standard deviation of the ratings assigned to all ratees on a particular performance dimension (Borman & Dunnette, 1975). Smaller standard deviations reflect greater range restriction. (Whether such range restriction represents leniency, severity, or central tendency can only be determined by examining the mean rating.) A second approach concentrates specifically on the proximity of the mean dimension ratings to the midpoint of the scale (DeCotiis, 1977; see, also, Korman, 1971, and other textbooks in industrial and organizational psychology). A third approach focuses on the degree of kurtosis, or peakedness, that characterizes frequency distributions of dimension ratings (Landy et al., 1976). A fourth approach is based on a Rater × Ratee × Dimension ANOVA in which the absence of a significant ratee main effect is interpreted as a lack of discriminability (Heneman, Schwab, Huett, & Ford, 1975).

Only the second of these four approaches concentrates specifically on central tendency, as previously defined. The remaining three operational definitions—standard deviations, kurtosis (leptokurtosis), and ANOVA—reflect only restriction of range and not necessarily central tendency.

Several other rating errors have been described in the recent rating literature, including "similar-to-me" error, "first and last impressions" error, and the "contrast effect" (Latham, Wexley, & Pursell, 1975). Since these criteria have played a relatively minor role in the study of rating behavior, however, they are not examined in any further detail in this article. Other more subjective criteria of rating quality, including raters' preferences for particular rating scale formats, as well as perceived relevance, difficulty, and accuracy of ratings (DeCotiis, 1977; Friedman & Cornelius, 1976), are similarly excluded from further discussion here. This is not meant to suggest that these other rating errors and subjective criteria are unimportant to the understanding and perhaps improvement of rating behavior. They are excluded

simply because the task of trying to eliminate some of the inconsistency surrounding halo, leniency and severity, and range restriction constitutes a more-than-sufficient challenge to the present authors.

Interrater Reliability or Agreement

In addition to the rating errors of halo, leniency and severity, and range restriction, the rating literature often uses a fourth criterion of rating quality—interrater reliability or agreement.[2] The justification for the use of this criterion is straightforward. Ratings are usually obtained primarily because other, more objective (countable), indices of behavior are unavailable. Thus, there are usually no more objective measures or criteria on which to rely for the purpose of validating the obtained ratings. The extent to which two or more raters independently provide similar ratings on given aspects of the same individuals' behaviors is therefore accepted as a form of consensual validity or convergent validity in the context of a multitrait–multirater matrix (Lawler, 1967).

Several authors, however, have expressed reservations regarding the status of interrater reliability or agreement as a criterion of rating quality. Buckner (1959) offered data suggesting that

> high agreement among the ratings assigned the same men by different raters does not necessarily imply predictable or valid ratings, and that disagreement among raters may be associated with predictability and possibly validity. (p. 60)

Freeberg (1969) concurred that "agreement between individuals . . . is not at all a necessary reflection of rater validity" (pp. 523–524) and concluded that his evidence supported Wherry's (Note 2) contention that "the reliability of a rating scale tells us very little about its value, since the apparent reliability may be due to bias rather than true score" (p. 39). Data reported by Borman (1975) suggest that successfully training raters to avoid halo error also produces lower interrater reliability and more accurate (valid?) performance profiles. Finally, Lumsden (1976), in a review of test theory literature, seriously (and sometimes not so seriously) questioned the use of reliability as a criterion of measurement quality and was extremely uncomplimentary toward the concept of a true score.

Five different operational definitions of interrater reliability or agreement emerged from the literature reviewed (Table 1). One approach examines the standard deviations of the ratings assigned to a particular ratee by several raters for a given (single) dimension of behavior (Bernardin, 1977). Smaller standard deviations for each of the individual behavior dimensions reflect greater interrater agreement. A second approach calculates correlations between pairs of raters who evaluate the same ratee on identical dimensions of behavior (Bernardin, Alvares, & Cranny, 1976). Larger correlations suggest greater interrater reliability, or convergent validity (Lawler, 1967). A third approach is based on the computation of intraclass correlation coefficients (Saal & Landy, 1977; see, also, Ebel, 1951; Shrout & Fleiss, 1979).

A fourth approach searches for a statistically significant ratee main effect, especially one that explains a sizable proportion of the rating variance, in a Rater × Ratee × Dimension ANOVA (Friedman & Cornelius, 1976). This approach, of course, is identical to one of the operational definitions of range restriction cited previously. The presence of a significant ratee main effect, then, can be interpreted as both interrater agreement and the absence of range restriction, whereas the absence of a significant ratee main effect can indicate both range restriction and a lack of interrater agreement. A fifth approach to quantifying interrater agreement checks for the absence of a significant Rater × Ratee interaction in a Rater × Ratee × Dimension ANOVA (Heneman et al., 1975). Since the presence of a statistically significant Rater × Ratee interaction has been interpreted as halo (Dickinson & Tice, 1977; Heneman et al., 1975), we have here a second example of a single statistical index that is used to operationally define two ostensibly different (at

[2] See Shrout and Fleiss (1979) for a discussion of the distinction between interrater reliability and agreement.

least at the conceptual level) psychometric properties of ratings.

Schmidt and Hunter (1977) contributed a sixth operational definition of interrater reliability to the literature. They stated that the only appropriate reliability coefficient for ratings used as criteria is the correlation across a reasonable time interval between ratings produced by different raters at Time 1 and Time 2. This strategy constitutes a specific modification of the correlational approach to interrater reliability described in the preceding paragraph.

Methodological Discrepancies

An additional source of potential confusion in the literature is the variety of different types of research designs, or data collection procedures, used by different researchers. These different designs are characterized by inherently limited capabilities for aggregating and analyzing data; and although apparently similar criterion indices can be derived, those indices naturally reflect such limitations.

The ideal data collection strategy is a complete Rater × Ratee × Dimension matrix, in which all of the raters evaluate all of the ratees on all of the behavior dimensions. Unfortunately, this strategy for collecting rating data is seldom feasible, especially outside of a rigorously controlled laboratory situation. The most commonly reported design is a partial Rater × Ratee × Dimension matrix, in which some of the raters evaluate some of the ratees on all of the dimensions (Borman & Dunnette, 1975; Cascio & Valenzi, 1977). (This design is sometimes reduced to a Ratee × Dimension design, summed or collapsed across raters). A third strategy for collecting rating data is a Rater × Dimension design, in which many raters evaluate the same ratee (Bernardin, 1977; Zedeck, Kafry, & Jacobs, 1976). A fourth procedure is the Ratee × Dimension design, in which each ratee is evaluated by a different rater. In a variation on this design, each rater evaluates several ratees, but none of the ratees are evaluated by more than a single rater. (See Cronbach, Gleser, Nanda, & Rajaratnam, 1972, for a more complete discussion of the effects of design on such studies.)

Except for the complete Rater × Ratee × Dimension design, each of these strategies yields data that are inherently limited regarding possible analyses. For example, data gathered according to a Rater × Dimension design, the usual approach in students' evaluations of their instructor's performance in the classroom, is simply incapable of yielding any Rater × Ratee interaction (an index of halo or interrater agreement) or any ratee main effect (an index of range restriction or interrater agreement).

In addition to using different data collection procedures, researchers assessing the quality of rating data may have one of two distinctly different purposes in mind. On the one hand, the research may be designed to estimate an absolute amount of rating error inherent in a particular rating system. This was typical of much of the earlier work (Jurgensen, 1950). On the other hand, the primary purpose of much of the more recent research (Bernardin, 1977; Friedman & Cornelius, 1976; Saal & Landy, 1977) has been to assess the relative psychometric strengths and weaknesses of two or more rating scale formats or rating systems. Depending on the purposes of the research, some of the operational definitions of rating quality, as well as some of the data collection procedures, will be more appropriate than others.

Implications of Different Operational Definitions of Rating Quality Criteria: Summary of an Empirical Study

Downey, Lahey, and Saal (Note 3) asked police sergeants to evaluate the job performance of their subordinates using two different rating scale formats (graphic rating scales and mixed standard rating scales) for the purpose of answering the following question: In the context of an empirical comparison of the psychometric qualities of ratings obtained using two different rating scale formats, can the use of particular operational definitions of the rating criteria affect the conclusions drawn?

Table 2
Summary of Rating Scale Format Comparisons Based on Different Operational Definitions of Rating Quality Criteria

Rating criterion	Operational definition	Psychometrically superior format[a]
Halo	Dimension intercorrelations	GRS
	Principal-component analysis	No difference
	Standard deviations	MSS[b]
	Rater × Ratee interaction	MSS
Leniency or severity	Mean dimension ratings	MSS
	Rater main effect	MSS
	Skewness	GRS
Central tendency or range restriction	Mean dimension ratings	GRS
	Standard deviations	GRS
	Kurtosis	No difference
	Ratee main effect	GRS
Interrater reliability or agreement[c]	Standard deviations	GRS
	Pearson product-moment correlations	GRS
	Intraclass correlations	GRS
	Ratee main effect	GRS
	Rater × Ratee interaction	MSS

Note. GRS = graphic rating scales; MSS = mixed standard rating scales. (Adapted from "Quantification of rating errors: Madness in our methods" by R. G. Downey, M. A. Lahey, & F. E. Saal, Note 3. Adapted by permission. Detailed analyses and interpretations appear in that report.)
[a] In some cases superiority was determined statistically, whereas in others more subjective (and perhaps practical) judgmental criteria were used.
[b] Substantial interrater differences emerged on this criterion.
[c] The general consistency that seems to characterize these findings evaporates when ratings on individual behavior dimensions are examined.

The results of this study are summarized in Table 2.[3] Different operational definitions of rating quality criteria did in fact suggest that conclusions regarding the relative psychometric superiority of graphic or mixed standard ratings are not entirely consistent. Particularly with respect to halo, leniency and severity, and interrater agreement, reliance on one operational definition produced results diametrically opposed to those that would have emerged with a different quantification strategy.

Since the choice of operational definitions of the psychometric qualities of rating data can actually determine the results of such a comparative study, that choice must be something more than an arbitrary exercise governed more or less by convenience. The final sections of this article facilitate conceptual and operational definition of criteria for gauging the psychometric properties of rating scale data and highlight the relative strengths and weaknesses of alternative data collection strategies (research designs).

Summary of Operational Definitions

To clarify the various types of rating errors and quality criteria that are found in the literature and discussed in the preceding sections of this article, we have standardized the notation. Table 3 consists of a Rater × Ratee matrix of hypothetical ratings on behavior dimension *d*. Each of the operational definitions of leniency and severity, halo, central tendency and range restriction, and interrater reliability and agreement is presented, using the notation of Table 3.

[3] Details of this study can be obtained on request from Ronald G. Downey or Frank E. Saal.

Table 3
Rater × Ratee Matrix for Dimension d

Ratee	Rater 1	2	3	···	j	Mean	Variance
1	X_{11d}	X_{12d}	X_{13d}	···	X_{1jd}	$M_{1.d}$	$S^2_{X1.d}$
2	X_{21d}	X_{22d}	X_{23d}	···	X_{2jd}	$M_{2.d}$	$S^2_{X2.d}$
3	X_{31d}	X_{32d}	X_{33d}	···	X_{3jd}	$M_{3.d}$	$S^2_{X3.d}$
⋮	⋮	⋮	⋮		⋮	⋮	⋮
i	X_{i1d}	X_{i2d}	X_{i3d}	···	X_{ijd}	$M_{i.d}$	$S^2_{Xi.d}$
Mean	$M_{.1d}$	$M_{.2d}$	$M_{.3d}$	···	$M_{.jd}$	$M_{..d}$	$S^2_{M.jd}$
Variance	$S^2_{X.1d}$	$S^2_{X.2d}$	$S^2_{X.3d}$	···	$S^2_{X.jd}$	$S^2_{Mi.d}$	$S^2_{X..d}$

Note. Subscript dot indicates summing over that factor.

Leniency and Severity

1. Mean dimension rating:

$$M_{..d} = M_{i.d} = M_{.jd}.$$

2. Rater main effect (for each dimension):

$$\frac{MS\ (\text{Raters})}{MS\ (\text{Raters} \times \text{Ratees})}.$$

3. Skewness (of dimension ratings): (a) skewness of all X_{ijd} scores (includes every rater–ratee combination), (b) skewness of the $X_{i.d}$ scores (focuses on ratees' ratings across raters), and (c) skewness of the $X_{.jd}$ scores (focuses on raters' ratings across ratees).

Halo

4. Dimension intercorrelations: $r_{X_{i.d},\ X_{i.d'}}$ (yields a $d \times d$ correlation matrix), where d' is another dimension from among the d total behavior dimensions.

5. Factor analysis, or principal-component analysis, of the $d \times d$ correlation matrix referred to in paragraph 4.

6. Rater × Ratee interaction:

$$\frac{MS\ (\text{Raters} \times \text{Ratees})}{MS\ (\text{Raters} \times \text{Ratees} \times \text{Dimensions})}.$$

7. Variance of the dimension ratings for each rater–ratee combination: $s^2_{Xij.}$.

Central Tendency and Range Restriction

8. Mean dimension rating: same as paragraph 1.

9. Absence of a ratee main effect (for each dimension):

$$\frac{MS\ (\text{Ratees})}{MS\ (\text{Raters} \times \text{Ratees})}.$$

10. Variance of ratings assigned to all ratees on single performance dimensions: (a) $s^2_{X..d}$, which includes all X_{ijd} scores; and (b) $s^2_{Mi.d}$, which focuses on ratees' average ratings across raters.

11. Kurtosis (of dimension ratings): (a) kurtosis of all X_{ijd} scores (includes every rater–ratee combination), (b) kurtosis of the $X_{i.d}$ scores (focuses on ratees' ratings across raters), and (c) kurtosis of the $X_{.jd}$ scores (focuses on raters' ratings across ratees).

Interrater Reliability and Agreement

12. Rater intercorrelations: $r_{X_{ijd},\ X_{ij'd}}$, for all possible combinations of raters, yielding a $j \times j$ correlation matrix. (Average correlations, based on Fisher's r to z transformation, over all possible rater combinations are often reported.)

13. Ratee main effect; see paragraph 9.

14. Absence of a Rater × Ratee interaction; see paragraph 6.

15. Intraclass correlations:

$$\frac{MS\ (\text{Ratees}) - MS\ (\text{Raters} \times \text{Ratees})}{MS\ (\text{Ratees}) + (j-i)\ MS\ (\text{Raters} \times \text{Ratees})},$$

which reflects the reliability of a single rater's ratings.

$$\frac{MS\ (\text{Ratees}) - MS\ (\text{Raters} \times \text{Ratees})}{MS\ (\text{Ratees})},$$

which reflects the reliability of the average of several raters' ratings.

16. Variance of ratings assigned to individual ratees on single performance or behavior dimensions: $s^2_{X_{i.d}}$.

Comment

Although the implications and constraints of these strategies for quantifying rating errors are examined in the next section of this article, a particularly salient observation is in order here. Textbooks in industrial and organizational psychology commonly discuss rating errors in terms of an individual rater's behaviors; however, all of the quantification methods previously described, save two (paragraphs 3c and 7), rely on an aggregate of several raters' data to reveal the amount of rating error present. The conceptual problems discussed earlier are not surprising, given this lack of congruency between conceptualization and quantification.

Typology of Rating Quality Criteria and Some Design Issues

Each of the various criteria of rating quality outlined in the previous section can be classified under one of three major headings: (a) *level* effects, based on mean values; (b) *dispersion* effects, based on standard deviation, variance, or kurtosis values; and (c) *multivariate* effects, based on measures of covariation for two or more behavior dimensions. Many of the difficulties associated with comparing the results of different methods for quantifying rating criteria are a function of the failure to carefully distinguish among the different kinds of information that the various methods afford.

For example, if we are concerned with the leniency or severity associated with a rating system, we can compute the overall (grand) mean of the ratings—a level effect. If the mean approximates the midpoint of the scale used, one may be tempted to conclude that no leniency or severity is present. Conversely, one might, with the same set of data, test for a rater main effect by examining the ratio of two variances (mean squares)—a dispersion effect—and conclude on the basis of a statistically significant F ratio that leniency or severity is present. Note that both outcomes can result from the same set of rating data.

If one subset of raters gives consistently high ratings, whereas another subset gives consistently low ratings, the overall effect is a grand mean at approximately the midpoint of the scale. Such relatively consistent subset behavior, however, would likely yield a statistically significant rater main effect. It would not be surprising if the luxury of such statistical significance persuaded the researcher to choose the rater main effect as the better overall index of the presence (or absence) of leniency or severity. Of course, a significant rater main effect does not necessarily preclude the possibility that all raters were using relatively high or low scale values. The simple fact is that these two indices reflect different properties of the rating data; one focuses on the average level of the ratings, whereas the other highlights the relative dispersion inherent in those ratings.

A comparable distinction between level and dispersion is central to the difference between central tendency and range restriction. The grand mean used to quantify leniency and severity is also descriptive of central tendency; a level effect is emphasized. Ratee main effects, on the other hand, reflect a concern with relative variances, or mean squares —a dispersion effect. Once again, information pertaining to both level and dispersion is necessary if valid conclusions are to be drawn.

Reliability, which can be understood as the degree of dispersion in ratee scores relative to the degree of dispersion in rater scores (over each rater), bears a conceptual and quantitative similarity to range restriction. (See the preceding section.) Reliability is often operationally defined in terms of an intraclass correlation; yet, Shrout and Fleiss (1979) demonstrated that there are at least six different forms of the intraclass correlation, depending on the assumptions in force. The issues raised by Shrout and Fleiss are particularly germane to this discussion, since they reflect a concern with the different types of analyses available, with the consequences of making different assumptions, and with

the implications of estimating reliability for one rater or for more than one rater. Once again, the decision to consider all or part of the available information is a crucial one.

Finally, halo is an index of multivariate effects across dimensions. Both the correlational and the factor (principal-component) analysis approaches seem to directly reflect this concept.

The Rater × Ratee × Dimension ANOVA approach, however, presents some logical (as well as analytical) problems. The major difficulty is that one must assume that exactly what one has stated is *not* true to conduct and interpret the analysis—that the dimensions are not conceptually independent but are merely different levels of the same factor or treatment variable. Ordinarily, raters evaluate several performance dimensions precisely because these dimensions are deemed to be qualitatively different aspects of performance or behavior. The ANOVA approach, however, implies that the different dimensions reflect a single construct or concept. The situation is analogous to treating height and weight as measures of the same dependent variable—perhaps size. Clearly, conceptual precision is not maximized.

The rating errors described are easily calculated when all of the raters have evaluated all of the ratees on all of the dimensions—a full design. Similar indices can be computed for partial designs, in which blocks of raters rate some but not all of the ratees on all the dimensions—ratees are nested under raters. ANOVA techniques may still be used (see Cronbach et al., 1972); a major difficulty, however, is insuring the comparability of the rater-ratee blocks, since such blocks (nests) can be confounded with other effects. The emergence of variance over each rater and each ratee renders the common methods for quantifying errors applicable.

What happens if we have neither a full nor a partial design, as when multiple raters evaluate a single ratee on several dimensions (e.g., students' ratings of their teacher)? The most popular procedures have involved (a) computation of dispersion values, either across raters (reliability) or across dimensions for each rater–ratee combination (halo); or (b) examination of a level effect (leniency or severity). The obvious weakness of these approaches is that they are incapable of separating ratee effects from rating error effects. If we turn first to the level effect, what are the implications of a high grand mean value across raters? Is this a case of blatant leniency, or are we merely dealing with an exceptional ratee? If two or more rating methods or systems are being compared, the same problem persists, since ratee effects are hopelessly confounded with method differences.

Examination of dispersion effects reveals similar difficulties. Little variance among a group of raters who evaluate a single ratee might indicate convergence of views (reliability), central tendency, leniency, severity, or an idiosyncratic quality of the ratee's behavior. Similarly, little variance across dimensions for a given rater–ratee combination is also ambiguous. If the ratee is a uniformly high- or low-performing individual, such a small variance value could point to the presence of halo when, in fact, the ratee is the consistent factor.

The most common methods for quantifying rating errors for single-ratee designs are seriously flawed. It appears therefore that research results based on such a design are potentially misleading and must be interpreted with extreme caution.

We have suggested that the quality of rating data can be assessed on the basis of three types of criteria—level effects, dispersion effects, and multivariate effects—and we have cautioned that unambiguous quantification of these effects requires multiple raters, ratees, and dimensions. The final section of this article describes a standard analytical approach to the quantification of these effects—a Rater × Ratee multivariate analysis of variance (MANOVA) using behavior dimensions as multiple dependent variables.

MANOVA Approach to the Typology of Rating Quality Criteria

The contradictions inherent in a Rater × Ratee × Dimension univariate ANOVA (see the previous discussion) suggest the potential usefulness of a Rater × Ratee MANOVA, us-

ing the behavior dimensions as multiple dependent variables, for quantifying the three major kinds of effects that comprise our typology of rating quality criteria. In addition to eliminating those contradictions, such an approach has the added benefit of simultaneously assessing level, dispersion, and multivariate effects.

One other departure from the traditional treatment of rating errors is an emphasis on the desirable characteristics of a rating system. Rating errors (the terminology itself is significant) have historically been defined in terms of particular undesirable rater behaviors, such as raters avoiding the midpoint of the scale (leniency or severity), restricting themselves to the midpoint of the scale (central tendency), or failing to discriminate among different aspects of behavior (halo). This tendency to attribute the outcome to rater behavior fails to properly acknowledge the relationships between and among raters, ratees, methods, traits, times, and so on, and their individual, collective, and interactive effects on the outcome (Landy & Farr, 1980).

For example, when a rating system produces a set of ratings that cluster about a point above the midpoint of the scale, we commonly label this leniency and attribute it to raters' misguided behaviors. Equally plausible explanations, however, are a rating scale that has a psychological midpoint that differs from the numerical midpoint (e.g., four negative anchors and only one positive anchor) or a group of ratees who are superior with respect to the dimension under consideration (e.g., strength for a group of weight lifters). The tendency to attribute this leniency to raters' misconduct tends to obscure these other possibilities. It is therefore advisable to define and emphasize the potentially desirable characteristics of a rating system; identification of the dynamics underlying a failure to achieve a certain characteristic should be a separate process. This approach reflects the ideas of Cronbach et al. (1972), who acknowledged the interdependence of each characteristic and the importance of interpreting the entire data set as a whole in their discussion of the theory of generalizability.

The following section summarizes the usefulness of a MANOVA approach for gauging the extent to which a set of rating data reflects desirable level, dispersion, and multivariate effects.

Level of Mean Rating

Rating systems should ideally yield mean values that approximate the midpoint of the rating scale, to maximize the potential rating variance. In addition, difficulties of interpretation arise when the effects of rating procedures, raters, ratees, or selected dimensions produce an overall (grand) mean that is substantially displaced from the numeric midpoint of the scale used. Any of the factors already mentioned, alone or in combination, can be responsible for such a displacement—a highly select group of ratees, lenient raters, or poor scaling procedures could produce a mean rating displaced in the positive direction.

Ratee Dispersion

Rating systems should adequately distinguish among ratees. Such discrimination would be reflected by the emergence of a significant ratee main effect. To facilitate comparisons among different research studies, intraclass correlations can be computed as an index of interrater reliability; this approach is analogous to a variance component analysis (see Shrout & Fleiss, 1979). In the traditional jargon, a significant ratee main effect suggests the absence of range restriction.

Rater Dispersion

Raters' average ratings, calculated over ratees, should be similar when their ratings are based on the same set or subset of ratees, which is indicated by the absence of a significant rater main effect. The results of different studies might be compared by calculating Pearson product-moment correlations for pairs of ratees and by obtaining average correlations through the use of Fisher's r to z transformation.

Multivariate Effect

Ratings should generally be assigned to a group of ratees so that those ratees are rank ordered differently on the different dimensions of behavior. In the MANOVA, such ratee differentiation would be reflected by a larger number of latent roots. In the traditional terminology, a greater number of latent roots suggests less halo. (All of these rating characteristics can be assessed in a similar manner, given a partial design in which some of the raters evaluate some of the ratees on all the dimensions. A Rater × Ratee × Group MANOVA, in which a group would consist of all raters who evaluate the same subset of ratees, would be appropriate. See Winer, 1971, p. 366, for the basic design.)

Concluding Remarks

Ellson and Ellson (1953) remarked that ratings have been with us for a long time. Considering the complexity of contemporary human behavior and the inability of objective measures to capture the richness of that behavior, there can be little doubt that ratings will continue to play a major role in both theoretical and applied psychological research for many years to come. It is therefore imperative that psychologists pursue research that is designed to maximize the desirable psychometric characteristics of ratings and to minimize or eliminate the undesirable characteristics.

Implicit in this charge to psychology is a demand for conceptual and operational clarity in the description of the various rating characteristics. No longer can we be "fuzzy" in our definition of leniency, for example, and then proceed to quantify the phenomenon with three different, noninterchangeable techniques. No longer can we define halo in terms of a particular rater's behavior and then proceed to quantify the phenomenon by aggregating data collected from a group of raters. No longer can we ignore the limitations with respect to aggregating and analyzing data that are inherent in some of the commonly used data collection designs (such as the single-ratee design).

The multivariate approach to evaluating the characteristics of ratings that is suggested in this article, along with the revised typology of rating characteristics, represents an attempt to circumvent some of the problems that have traditionally plagued rating research. It is hoped that the increased precision at both the conceptual and operational levels that is available through this approach will facilitate psychology's quest for a better understanding of the complex phenomena of rating behavior. Failure to pursue alternative approaches such as this will surely sentence rating research to many more years of inconsistency and confusion.

Reference Notes

1. Downey, R. G., & Saal, F. E. *Evaluating human judgment techniques.* Paper presented at the meeting of the American Psychological Association, Toronto, Ontario, Canada, August 1978.
2. Wherry, R. J. *The control of bias in ratings: VII. A theory of rating* (Final Report No. 922). Department of the Army, Personnel Research Branch (now Army Research Institute for the Behavioral and Social Sciences), February 1952.
3. Downey, R. G., Lahey, M. A., & Saal, F. E. *Quantification of rating errors: Madness in our methods.* Manuscript submitted for publication, 1980.

References

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

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.

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

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. Consistency of rating accuracy and rating errors in the judgment of human perform-

ance. *Organizational Behavior and Human Performance,* 1977, *20,* 238–252.

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

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

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

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.

Dickinson, T. L., & Tice, T. E. The discriminant validity of scales developed by retranslation. *Personnel Psychology,* 1977, *30,* 255–268.

Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika,* 1951, *16,* 407–424.

Ellson, D. G., & Ellson, E. C. Historical note on the rating scale. *Psychological Bulletin,* 1953, *50,* 383–384.

Finley, D. M., Osburn, H. G., Dubin, J. A., & Jeanneret, P. R. Behaviorally based rating scales: Effects of specific anchors and disguised scale continua. *Personnel Psychology,* 1977, *30,* 659–669.

Ford, A. Neutralizing inequalities in ratings. *Personnel Journal,* 1931, *9,* 466–469.

Frandsen, A. N. *Educational psychology* (2nd ed.). New York: McGraw-Hill, 1967.

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.

Garrett, H. E., & Schneck, M. R. *Psychological tests, methods, and results.* New York: Harper, 1933.

Guilford, J. P. *Psychometric methods.* New York: McGraw-Hill, 1954.

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

Heneman, H. G., III, Schwab, D. P., Huett, D. L., & Ford, J. J. Interviewer validity as a function of interview structure, biographical data, and interview order. *Journal of Applied Psychology,* 1975, *60,* 748–753.

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,* 130–134.

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

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

Kingsbury, F. A. Analyzing ratings and training raters. *Journal of Personnel Research,* 1922, *1,* 377–383.

Kneeland, N. That lenient tendency in rating. *Personnel Journal,* 1929, *7,* 356–366.

Korman, A. K. *Industrial and organizational psychology.* Englewood Cliffs, N. J.: Prentice-Hall, 1971.

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

Landy, F. J., & Farr, J. L. Performance rating. *Psychological Bulletin,* 1980, *87,* 72–107.

Landy, F. J., Farr, J. L., Saal, F. E., & Freytag, W. R. Behaviorally anchored scales for rating the performance of police officers. *Journal of Applied Psychology,* 1976, *61,* 750–758.

Landy, F. J., & Trumbo, D. A. *The psychology of work behavior.* Homewood, Ill.: Dorsey Press, 1976.

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

Lumsden, J. Test theory. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 27). Palo Alto, Calif.: Annual Reviews, 1976.

McCormick, E. J., & Tiffin, J. *Industrial psychology* (6th ed.). Englewood Cliffs, N. J.: Prentice-Hall, 1974.

Motowidlo, S. J., & Borman, W. C. Behaviorally anchored scales for measuring morale in military units. *Journal of Applied Psychology,* 1977, *62,* 177–183.

Murray, H. A. *Explorations in personality.* New York: Oxford University Press, 1938.

Newcomb, T. An experiment designed to test the validity of a rating technique. *Journal of Educational Psychology,* 1931, *22,* 279–288.

Nisbett, R. E., & Wilson, J. D. The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology,* 1977, *35,* 250–256.

Phares, E. J. *Clinical psychology: Concepts, methods, and profession.* Homewood, Ill.: Dorsey Press, 1979.

Rizzo, W. A., & Frank, F. D. Influence of irrelevant cues and alternate forms of graphic rating scales on the halo effect. *Personnel Psychology,* 1977, *30,* 405–417.

Rugg, H. Is the rating of human character practicable? *Journal of Educational Psychology,* 1921, *12,* 425–438. (a)

Rugg, H. Is the rating of human character practicable? *Journal of Educational Psychology,* 1921, *12,* 485–501. (b)

Rugg, H. Is the rating of human character practicable? *Journal of Educational Psychology,* 1922, *13,* 30–42. (a)

Rugg, H. Is the rating of human character practicable? *Journal of Educational Psychology,* 1922, *13,* 81–93. (b)

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

Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology,* 1977, *62,* 529–540.

Schneider, D. J. Implicit personality theory: A review. *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.

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

Shrout, P. E., & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin,* 1979, *86,* 420–428.

Stanley, J. C. Analysis of unreplicated three-way classification with applications to rater bias and trait independence. *Psychometrika,* 1961, *26,* 205–219.

Stockford, L., & Bissell, H. W. Factors involved in establishing a merit-rating scale. *Personnel,* 1949, *26,* 94–116.

Thorndike, E. L. A constant error in psychological ratings. *Journal of Applied Psychology,* 1920, *4,* 25–29.

Webb, E. Character and intelligence. *British Journal of Psychology Monograph Supplement,* 1915, *1* (No. 3).

Wells, F. L. A statistical study of literary merit. *Archives of Psychology,* 1909 (No. 7). (Monograph)

Wexley, K. N., & Yukl, G. A. *Organizational behavior and personnel psychology.* Homewood, Ill.: Irwin, 1977.

Willingham, W. W., & Jones, M. B. On the identification of halo through ANOV. *Educational and Psychological Measurement,* 1958, *18,* 403–407.

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Zedeck, S., & Blood, M. R. *Foundations of behavioral science research in organizations.* Monterey, Calif.: Brooks/Cole, 1974.

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

Received October 1, 1979 ■