APA CENTENNIAL FEATURE

The Criterion Problem: 1917–1992

James T. Austin
Ohio State University

Peter Villanova
Northern Illinois University

Individuals differ on multiple aspects of their job-role behavior; criteria are measures that attempt to capture these differences. Measures of criteria are used by several constituencies within applied psychology. Among them, researchers used criteria for the evaluation of theories of work behavior, the effective administration of human resources and the provision of feedback to individuals. One index of the importance of criteria is the observation that most, if not all, of the pioneers of industrial-organizational psychology addressed this issue during their careers. This article reviews conceptual and methodological developments pertaining to the criterion problem since 1917, using as an organizing device dimensions, methods of measurement and analysis, and categorizing frameworks. A shift away from an emphasis on brute prediction toward a balanced treatment of both empirical and conceptual issues is highlighted by calls for the validation of criteria and by increased attention to modeling performance, as well as a recognition of multiple perspectives and competing values from which to view performance and criterion measurement.

People in organizations behave according to their role perceptions and the roles assigned to them by others. The resulting behaviors and associated outputs, which are often limited by situations (Viteles, 1925–1926b) and may vary widely across individuals (Tiffin, 1942; Viteles, 1932), form a large part of the subject matter of industrial–organizational (I-O) psychology. Individual variability on criterion measures interests researchers and managers, who attempt to measure and influence it. Because criteria are essential for evaluation of individuals, programs, and organizational interventions (Schmitt & Klimoski, 1991), the practical significance of criterion measurement has been recognized since the beginning of industrial psychology (Münsterberg, 1913; Scott, 1917). However, the systematic study of criteria per se to expand scientific knowledge about work behavior is a fairly recent phenomenon. This statement does not imply that criterion research has not been pursued vigorously, but rather that the necessary conceptual, taxonomic, and methodological prerequisites for the pursuit of understanding criteria were relatively undeveloped in the first half of the 20th century.

The legacy of the first 60 years of scientific research on criteria (roughly from 1917 through the appearance of the I-O handbook in 1976) can be conveyed through the appearance of the term *criterion problem* (e.g., Flanagan, 1956; Landy & Farr, 1983; Smith, 1976). The term is often invoked to admonish students and sponsors of research about the difficulties involved in the process of conceptualizing and measuring performance constructs that are multidimensional and appropriate for different purposes. The criterion problem endures for additional reasons that we review below. Unlike predictor constructs (e.g., general cognitive ability), criterion constructs often require additional translations between concepts and measurement operations. Binning and Barrett (1989), for one, noted that the translation from behavior to results may be constrained by situational factors.

In addition, the choice of dimensions to represent a criterion construct depends on how broadly one initially construes the conceptual criterion (Nagle, 1953; Toops, 1944). Further difficulties arise if we also accept Bailey's (1983) assertion that dimensions of criteria are context dependent, in which case the same measure used to represent a conceptual criterion in one setting may prove unsuitable in a different context. Finally, the failure to articulate the values involved in decisions to include some measures of performance as criteria while excluding others makes the criterion problem that much more oblique (cf. Fiske, 1951; Villanova, 1992).

This article is dedicated to Robert L. Thorndike, 1910–1990, who made monumental contributions to the study of criteria during his long career at Teachers College, Columbia University. Both authors contributed equally to the preparation of this article.

We thank the following individuals, who critically and constructively commented on earlier versions of this article: Walter Borman, Ray Katzell, Lloyd G. Humphreys, Ann-Marie Ryan, Robert Heneman, Kristin Boyle, David Hofmann, and Patricia Cain Smith. We also thank two anonymous reviewers who were extremely helpful; Robert Guion, who suggested the inclusion of a section on semantic usage of the term *criterion*, as well as the organizing framework of dimensions, methods, and categorization schemes; H. John Bernardin for discussions of the criterion problem that improved the article; and Tammy Gibson and Meg Murray, who proofread earlier versions. We alone are responsible for the content of the article.

Correspondence concerning this article should be addressed to James T. Austin, Department of Psychology, Ohio State University, 142 Townshend Hall, 1885 Neil Avenue, Columbus, Ohio 43210-1222.

Journal of Applied Psychology, 1992, Vol. 77, No. 6, 836–874

Organization and Purpose

This article reviews conceptualizations, technical advances, and controversies in the measurement and use of criteria since the formal beginnings of the discipline. We begin the review in 1917 and proceed in 20-year intervals up to the present. The focus will be on evaluation of conceptual treatments of criteria, using a three-part framework, but there is certainly no scarcity of empirical studies (evaluated in *Annual Review of Psychology* chapters since 1950; Borman, 1991; Landy & Farr, 1980, 1983; Ronan & Prien, 1966, 1971; Smith, 1976). However, with few exceptions (Bailey, 1983; James, 1973; Smith, 1976), efforts to integrate conceptual and empirical advances in criterion research to further understand criterion issues are mostly products of recent history (Campbell, 1990a).

Aside from intellectual curiosity, there are several reasons to examine the historical record of applied psychology (Darley, 1968; Guion, 1976). Among them is the observation that many current concerns in criterion measurement were noted very early (Bingham, 1926; Burtt, 1926; Hull, 1928; Kornhauser, 1926–1927; Strong, 1918; E. Thorndike, 1918; Viteles, 1925–1926b, 1932). Contamination and deficiency as bias were known in those days; different perspectives of individuals and institutions were also recognized. As one specific example, Symonds (1931, pp. 112–113) proposed and used partial correlation techniques to estimate the size of the halo effect. His suggestion preceded by some 50 years the application of the same procedure by Landy, Vance, Barnes-Farrell, and Steele (1980). Such major problems, many of which are not yet solved (and, indeed, may be insoluble), attest to the importance of the history of criterion measurement. A second important reason is the contribution of criteria to productivity at multiple levels of analysis, a current concern with national implications (Campbell & Campbell, 1988; Zeidner, 1987). This issue focuses attention on the possibility that measures of performance can be expressed and evaluated at individual, group, and organizational levels in similar ways (differences would suggest a requirement for composition theories; cf. Dansereau & Markham, 1987; Rousseau, 1985). A final reason is that the problem of criteria, defined as selection of dependent variables, is not unique to I-O psychology. It stretches throughout applied and basic psychology and into other social sciences. Before beginning the review, however, we discuss the organizing framework and semantic usages of the term criterion as one way to trace historical developments.

The organizing framework of this review applies three facets to tracing the history of criterion measurement. The first focuses on how the dimensions of criteria have been conceptualized and studied. This facet includes characterizations of the scope of an activity domain and its underlying determinants (e.g., quantity of production, tenure, ability, effort, constraints). Methods of measurement and analysis, the second facet, involves discussion of development, measurement, and analyses of different criteria (e.g., production counts, ratings, personnel records). The third facet, categorizing frameworks, is concerned with conceptual attempts to group criteria (e.g., time, type, and level as proposed by Weitz, 1961). Classification is an essential part of science, as noted by Pearson in his *The Grammar of Science* (1892/1957) and more recently by Fleishman and Quaintance (1984) and Cattell (1988b). We are not claiming that the facets are independent, in fact, it seems to us that there may be some interdependence.

Within each time interval, the framework is applied to criterion measurement to provide an understanding of progress and problems. We also include brief annotations of studies to illustrate the modal criteria of choice for validation and their treatment during each period. Our selection of specific studies was not meant to imply that they were more revolutionary or more significant than those not reviewed. In fact, like many scientific advances, they represent incremental gains in knowledge.

Semantics of the Term *Criterion*

There are many ways in which the term *criterion* has been used (Austin, Villanova, Kane, & Bernardin, 1991; Guion, 1991). Scrutiny of several dictionaries of psychological terms revealed multiple usages of the term. Warren (1934) defined the term in two ways. It was either a "standard adopted for qualitative comparison" (p. 64) or a "test of truth or a basis for judgment" (p. 64). He also distinguished a "criterion" from a "standard" of measurement by distinguishing between qualitative (criterion) and quantitative (standard). Presumably, the judgment refers to an evaluation of some other measure(s). English and English (1958) provided four definitions 24 years later. Their definitions were, first, a "basis for a judgment or qualitative comparison"; second, a "behavior goal by which progress is judged"; third, a "variable, comparison with which constitutes a measure of validity"; and fourth, a "dependent variable, the variable to be predicted" (p. 130). They also distinguished criterion from standard, but in a different way. Specifically, criterion referred to an external basis for judgment, and standard represented an internal and quantitative comparison. Reber (1985) defined criterion as "a standard against which a judgment, evaluation, or classification can be made" (p. 166). He provided several alternatives, including use as a learning standard, as an internal threshold in signal detection theory, and as a term in statistics for the variable to be predicted.

There was also definitional multiplicity within the applied domain. Though the specific term criterion did not appear in his early writings, Scott (1917) was among the first to discuss validation in the sense that test scores should correlate with some external measure. He outlined four procedures for validating scores for vocational selection, including what are now defined as concurrent and predictive strategies. Scott also described two criterion-group strategies that consisted of comparing the test score distributions of different groups (successful vs. unsuccessful incumbents, or "ringers" vs. applicants and successful incumbents, or "experts" vs. applicants). A difference between the two group comparison strategies was that the ringers approach used both high- and low-performing groups as standards of comparison (with group membership defined by the organization).

Hull (1928) used criterion to refer to the aptitude represented by scores on a test. For him the criterion was not a dependent variable, but more an alternative measure of the same concept a test was designed to measure, akin to the contemporary definition of convergent validity. Bechtoldt (1947) defined a criterion as a standard for evaluating other measures and as a means of

describing individual performance on a success continuum (p. 357). Success was recognized as nearly always multidimensional in character, meaning that its sources of variance were complex. Consistent with prevailing opinion in the 1940s and 1950s, Brogden and Taylor (1950a) preferred a unitary criterion (a single variable or weighted linear combination) expressed on a monetary scale to represent an employee's value to the organization. Starting in the 1940s and through the 1960s up to the present, factor analyses repeatedly showed the multidimensionality of criteria (e.g., Bolanovich, 1946; Campbell, McHenry, & Wise, 1990; Hulin, 1963; Roach & Wherry, 1970; Rush, 1953). Additional logical arguments against unitary criterion measures (Dunnette, 1963a; Ghiselli, 1956; Guion, 1961) focused on the inadequacy of single measures to represent complex phenomena, although 85% of the reports in the Validity Information Exchange (VIE) used only a single criterion (Lent, Aurbach, & Levin, 1971). The need to investigate criterion measures was argued by some (Bechtoldt, 1947) and disputed by others (Astin, 1964; Ronan & Prien, 1966). Ronan and Prien (1966) defined a criterion as a "measurement of the manifestations of performance behavior based upon characteristics of individuals as they affect and are affected by situational and organizational characteristics" (p. 4). According to Guion (1976), the controversy between single and multiple criterion advocates (reviewed by Schmidt & Kaplan, 1971) had more to do with an evolution in validity concepts that was taking place during this period (e.g., American Psychological Association [APA] 1954; Cronbach, 1971; Cureton, 1951; Messick, 1989; R. L. Thorndike, 1971) than with substantive disagreements among scholars as to the appropriate rules of inference. Recently, attempts to model a "criterion space" have reflected progress, sometimes unequally, in conceptual and statistical domains. Guion (1991) pointed out that validation requires specification of both predictor and criterion constructs (cf. Landy, 1986). He differentiated between psychometric and job-relatedness usages of the term criterion. The two uses are not always mutually exclusive, although a theoretical–practical distinction between them can be used to classify purposes of measurement. Wiley (1991) made a similar distinction between test and construct validation based on the focus of the study.

Bingham (1926) was perhaps the first to use the word criterion in one of the two ways that it is frequently used today, as "something which may be used as a measuring stick for gauging a worker's relative success or failure" (p. 1). A second contemporary usage of the term criterion to mean that which is predicted from test scores but that is nonredundant with the concept represented by test scores can be traced back to Burtt (1926). He defined a criterion as "an index of occupational proficiency which is used in evaluating the tests designed to predict that proficiency" (p. 169). Although not explicit in those writings, this represents a causal usage in that the knowledges, skills, abilities, and orientations (KSAOs) assessed by predictor measures are believed to underlie and support job success. At a minimum, predictors are hypothesized to covary with job success.

Current use of the term criterion to mean both the sample of the performance domain to be predicted and the level of performance considered acceptable (or a standard) can be traced back to Weitz (1961), who maintained that both decisions were part and parcel of the criterion development process. Weitz's definition of criteria implied that criteria were distinguishable from performance with performance being a more inclusive concept. However, Weitz's discriminative use of the term has not always been heeded by scholars. As one example, Schneier and Beatty (1979) defined performance as behavior that is evaluated (p. 4). That definition of performance is difficult to reconcile with Weitz's treatment of the term given that it suggests that performance and criteria share equal status. We disagree with this implied subordination of the word criterion. We (Austin et al., 1991), like Bailey (1983) and Weitz (1961), would contend that criteria occupy a special status as a function of being critical samples of the more extensive performance domain. They are critical in the sense of their value to multiple sets of potential users (e.g., researchers, managers and members of organizations, and individuals who evaluate social programs).

A value orientation implies that the choice of criterion measures involves, however indirectly, the interests of multiple constituencies and may not exclusively reflect rational economic or scientific concerns (Cronbach, 1988; Fiske, 1951; Messick, 1989). Cronshaw (1989) applied similar logic to criteria and wondered whether science and value could be merged in a coherent manner. For instance, user constituencies in organizations may value utilization information about rater and ratee perceptions of the fairness of performance appraisal procedures as well as psychometric information (Jacobs, Khafry, & Zedeck, 1980). Alternatively, researchers may value information about the construct validity of the criteria used to represent job performance and so may include measures of job performance that are not essential from other perspectives. Similar arguments have been presented in opposition to goal-based conceptions of organizational effectiveness. Zammuto (1984) and Keeley (1984) have independently developed a similar perspective on organizational effectiveness, which constitutes a criterion conceptualization and measurement problem at a higher level. Briefly, constituencies have the characteristics of political coalitions, including different perspectives and sources of power that interact over time. Zammuto (1984) emphasized that organizational effectiveness is constrained by values of constituencies and time (as well as their interaction); Keeley (1984) analyzed "participant-interest" theories and concluded that a harm-minimization rule might be an appropriate compromise in certain cases.

To avoid confusion as to how we use the term below, we offer the following by way of a working definition: *A criterion is a sample of performance (including behavior and outcomes), measured directly or indirectly, perceived to be of value to organizational constituencies for facilitating decisions about predictors or programs.* This working definition restricts our discussion to instances in which a criterion is conceptualized as meaning something different from the concept represented by predictor scores though such scores may correlate with scores on a criterion measure. By this we mean that criterion scores represent in part the causal effects of individual differences in predictor scores but are conceptually distinct from the construct reflected in the predictor scores (Hull, 1928).

Precursors

Just before the development of industrial psychology, an emphasis on criteria was provided by the adherents of scientific

management, who focused on the measurement of work outputs for work simplification and increased output. The governing variables of interest were output quantity and quality. The primary methods were observation; motion and time studies of worker behavior; and the systematic study of materials, tools, and procedures. Time study was advocated by Frederick W. Taylor (1911/1947), and motion study was advocated by the Gilbreths (Gilbreth, 1909, 1911; Gilbreth & Gilbreth, 1916). Their taxonomy of work movements included 17 *therbligs* (Meredith, 1953).

Viteles (1932), as well as other psychologists (e.g., Farmer, 1921), extensively critiqued the contributions of scientific management to early industrial psychologists. The gist of their critiques was that scientific management practitioners were too concerned with job-specific production measures. Relative emphasis on the time or motion aspects of work study was disputed (cf. Barnes, 1940; Karger & Bayha, 1987). Such efforts were characterized by an imperfect blend of psychological and engineering approaches that used inductive strategies to control worker behavior. Such studies were driven by the pragmatic interests of industry and emphasized the viewpoint of management, which valued those efforts for the solutions they posed for specific problems in unique situations. Because of those characteristics and the nonsystematic application of the scientific method, they are more properly classified as precursors to the industrial psychology of the following decade. This is not to say that early industrial psychologists could not be described by many of the same characteristics (Sokal, 1984). However, the more systematic application of the scientific method by psychologists acted to correct some early mistakes, whereas scientific management evolved more toward an engineering orientation. Gilbreth and Gilbreth (1924) themselves noted several differences between the perspectives of industrial psychologists and engineers. Such reasons are why we have chosen to focus on the efforts of psychological researchers, while acknowledging the propaedeutic contributions of scientific management (Locke, 1982).

1917–1939

We begin the review with the year 1917, which marked the founding of the *Journal of Applied Psychology* (Geissler, 1917) and the entry of the United States into World War I. Before proceeding with the framework, we offer several general observations to set the context of the times. Applied psychologists of those times were trained in experimental psychology (often in Germany) and occupied academic positions. In the first two decades of this century, several early members of the APA were interested in individual differences and the application of psychology to everyday life (e.g., Walter V. Bingham, James McK. Cattell, Hugo Münsterberg, Walter Dill Scott, and E. L. Thorndike, as described by Landy, 1992). Industrial, educational, and clinical domains were prominent areas for application, and all flowered at roughly the same time. However, not all psychologists enthusiastically embraced the initial editorial of the *Journal of Applied Psychology* that urged colleagues, through their research, to "contribute their quota to the sum-total of human happiness" (Hall, Baird, & Geissler, 1917, p. 6). In fact, Yerkes' (1917) first contribution to the *Journal of Applied Psychology* criticized Terman's 1916 Stanford Revision of the Binet Scale on the grounds that though it was "technologically useful," it possessed "little research value" (p. 115) compared with Yerkes own point scale method! Hollingworth and Poffenberger (1924) cautioned that one of the great dangers for applied psychology was that "too much may be expected of it, and that it may be extended into fields where it is not prepared to go" (p. 18).

Functionalism and pragmatism were dominant themes of the respective American psychological and philosophical climates (Buxton, 1985). There were also the beginnings of a shift toward behaviorism that represented an interest in the results of conscious process, rather than the processes themselves (Angell, 1911). Together those intellectual themes converged on the idea that the facts to be explained had to be observable, functional, and significant to human existence.

Personnel selection issues dominated industrial psychology at its inception and for many subsequent years (Katzell & Austin, in press; Scott, 1920). The population of major interest was lower level employees (e.g., clerical, transportation, and production workers), who, because of the transition from an agricultural to a manufacturing economy, worked for organizations embedded in an increasingly complex industrial society consisting of new mass production technology, urban living conditions, and rising immigration (Allen, 1931; Braeman, Bremner, & Brody, 1968; DeSantis, 1989; Wiebe, 1967). Reacting to a quest for "order," a term used by Wiebe (1967), mental tests and criterion measures promised classification and thus provided some sense of order for both employers and employees. The early state of the art in personnel psychology was described by Link (1919, 1920), Kornhauser (1922), Kornhauser and Kingsbury (1923), Freyd (1923–1924), and Kelley (1919). Those writers agreed in principle on job analysis (Strong & Uhrbrock, 1923), the use of tests as well as other predictors (Hull, 1928), and validation against criteria of work success (Bingham, 1926). The criterion thus served as a standard for the evaluation of predictors, usually tests, although it was realized that the methods could be applied to any device used to make personnel decisions (cf. Kornhauser & Kingsbury, 1923, p. 43). Guion (1976, 1991) provided a modern perspective on selection during that period, although he pointed out that Freyd's (1923–1924) original 10 steps (or, we note, Kornhauser & Kingsbury's 6 steps) for establishing the criterion-related validity of predictors have not changed all that much. The apparent successes of intelligence testing in the U.S. Army (Hull, 1928; Link, 1919; Strong, 1918; Yoakum & Yerkes, 1920), as reviewed by Hale (1982), coupled with the beginnings of vocational guidance (Hollingworth, 1920; Münsterberg, 1913; Parsons, 1909), held the promise that people could be matched effectively to jobs, using tests and other devices that were capable of evaluation (Scott, 1917).

There are several counterpoints to this positive picture. First, Sokal's (1987) edited book contains suggestions that there were squabbles among the early researchers (especially between Scott and Robert M. Yerkes) that may have impaired the contribution of psychology to the war effort (von Mayrhauser, 1987, 1992). Also, as might be expected at the formation of a disciplinary area, disputes about the definition and even the possibility of applied psychology were common (e.g., Bingham, 1923; Freyd, 1926; Geissler, 1917; Roback, 1917). Writing else-

where, Sokal (1984) concluded that the pattern for scientific psychology during the subsequent decade of the 1920s consisted of limited success, overconfidence and overstatement, and retrenchment. The oft-cited example of intelligence tests provides the best illustration of this cycle (Hothersall, 1990, gives a useful description of this phase of American psychology). Second, Guion (1991) recently observed that tests are often the only component of selection systems held to a standard of validation, an observation that applied equally to selection systems of this early era (Bergen, 1934–1935). Other instruments such as interviews, application forms, and experience often entered into selection decisions then (Bills, 1938–1939) and may relate differentially to criteria.

Dimensions

How have criteria been conceptualized in terms of their dimensionality? Dimensionality within each interval was studied by several strategies. First, we examined the prominent writings of the period to ascertain listings and analyses of various criterion measures. Second, we examined published studies to ascertain their use of criterion measures. We note that the method of choice during this period was rational listing of criterion measures, because there were no quantitative methods for the empirical determination of dimensionality. Spearman's methods (antecedents of factor analysis), although growing in popularity, were applied mainly to ability measures. Writings by Kornhauser and Kingsbury (1923), Bingham and Freyd (1926; Bingham, 1926), Burtt (1926), and Viteles (1925–1926b, 1932) and a symposium on criteria (Hoppock, 1936) were chosen for the former task. Those discussions revealed substantial convergence and broad coverage of the criterion space.

Kornhauser and Kingsbury (1923) discussed output, ratings, and length of service classes of measures. They preferred output measures, but recognized that such measures were not always available when evaluating tests, and maintained that length of service should always be a subsidiary criterion measure. This was despite the problems that industry was experiencing with turnover at the time (Bezanson, Willits, Chalufour, & White, 1922). Bingham and Freyd (1926) listed and discussed the following measures: training time, training performance, quantity, and quality of output (recommended as the best measure, but only if uncontaminated), work samples, accidents, salary, commissions and bonuses, length of service, advancement, degree of responsibility, membership in professional societies, trade status, and ratings. Burtt (1926) presented output and supervisory ratings as the major types of measures (work samples were added in his 1942 revision). Under miscellaneous criteria Burtt presented quality of work, amount of preliminary training, length of service, and advancement. He too preferred production measures and provided 17 specific variants for retail selling. He recommended that motivation, units of measurement, absence, experience, and reliability be equated for meaningful validation when using production measures. Viteles (1932) gave a similar list, divided into 12 objective (e.g., quantity of output, number of accidents, rate of advancement) and two subjective (graphic rating scales and ranking methods) classes. Viteles stressed that criterion choice was dictated by the job and by the goals of the researcher; he also noted that different criteria might be better suited for different predictors. Viteles (1925–1926b), in an earlier article, had been critical of personnel managers and researchers for the poor criteria generally found in practice and research. He also urged adding clinical-dynamic approaches to the statistical approach of personnel psychology (Viteles, 1925, 1936). This could have reflected the influence of his advisor, Lightner Witmer, a pioneer in the delivery of clinical treatments (Viteles, 1931), and of Pierre Janet, whom he met while on a postdoctoral fellowship in Paris (Viteles, 1974).

A symposium sponsored by the National Occupational Conference (Hoppock, 1936) contained diverse presentations on criteria. The sections consisted of treatments of the problem as a whole, arguments for five specific criterion measures, criteria recommended for vocational guidance work, an argument by Viteles in favor of a clinical-dynamic criterion, and an annotated bibliography. A highlight was Kurt Lewin's discussion of success and failure in terms of aspiration levels, self-esteem, and the psychological situation. Goal attainment was an engine-driving satisfaction, which was just then becoming a researched topic. The breadth of measures discussed in this symposium was greater than that used by the practitioner community.

The practice of empirical validation dated from the work of Binet and Simon (1908), advocates of "testing the tests" before applying them (cf. Goodenough, 1949), even though researchers seemed to prefer concurrent and convergent strategies in which multiple measures (intelligence tests) were administered and intercorrelated (cf. Jordan, 1923). Scott (1917), as noted earlier, described several methods of evaluating vocational selection, differentiating it from vocational guidance by its focus on the employer rather than the employee, and noting especially its ease of validation relative to guidance. We chose to examine three representative studies, one of shell inspectors (Link, 1918), one of metalworkers (Pond, 1927), and one of a sample of eighth graders tested and followed by E. L. Thorndike and associates (E. L. Thorndike et al., 1934).

Link (1918) studied two jobs, involving inspection and gauging of cartridges, held by women at the Winchester Company. He selected eight tests of sensory and specific abilities (e.g., eyesight, cancellation, and number checking) and correlated them with average production in pounds over a 4-week period using concurrent validation. His single criterion of output (over 4 weeks) was not evaluated psychometrically, although it was averaged to increase reliability and standardized on only one type of shell. The results indicated different patterns of rank correlation for the two different jobs. Differences in correlational patterns with the quantity criterion were explained by differences between the jobs. A crude cross-validation was conducted by applying the three best tests to several new groups of inspectors, and the results indicated confirmation for three of four groups. Link's (1919) book, published the following year, presented those results and others (along with making the tests available through a publisher) to provide a systematic elaboration of the work of Münsterberg (1913).

Pond (1927), as part of her dissertation at Yale University, evaluated the use of eight ability tests for the selection of 29 job classes of factory operatives at the Scoville Corporation, using a predictive design in which predictor scores were not used in selection or placement. Her criteria included highest weekly

pay, increase in earnings, supervisor's ratings, and terminations. Again, there was little analysis of the psychometric properties of the criterion measures, although Pond stated that the intercorrelations among the four criterion measures were low. Her results indicated that test scores and criteria were nearly uncorrelated, but critical scores were determined for roughly one third of the occupational groups (9 of 29) separately for termination and supervisory rating criteria.

E. L. Thorndike et al. (1934) evaluated the predictive efficacy of vocational tests applied to students in New York city schools (Lorge, 1933, 1936). Approximately 2,225 students were assessed in eighth grade (1921–1922) with a battery of predictors including general, clerical, and mechanical ability tests. They were followed up in school (grades and completion) and also some eight years later with several vocational criteria. Although nine potential criterion measures were defined in the report, the researchers actually used three: average annual earnings across jobs held, a weighted average of the levels of those jobs, and a weighted average of satisfaction and interest in their jobs. Those criteria were analyzed separately. The jobs reported by the subjects were categorized into four groups according to whether they appeared to be mechanical, clerical, some combination of the two, or professional. The results were interpreted as disappointing by E. L. Thorndike, who stated that guidance was similar to "billiards" (Lorge, 1936).

Very quickly, however, criticisms of this study were advanced from several quarters (Husband, 1936; Macrae, 1934; Paterson, 1935; Viteles, 1936). Perhaps the most serious flaw was pointed out by Husband (1936), who noted that there was no attempt to implement any formal guidance, such as counseling the participants, on the basis of their test scores and backgrounds. In particular, at that time vocational guidance was defined by the National Vocational Guidance Association as assisting individuals to choose an occupation, prepare for and enter it, and succeed in it. Yoakum (1922–1923) discussed the mechanics of vocational guidance by elaborating the three-part framework of Parsons (1909), which consisted of assessing the person, the job, and the match between them. The gist of the criticism was that E. L. Thorndike and associates were using the results of a study of the predictive power of tests (general and specific abilities) to conclude that vocational guidance was flawed or that they generalized too widely. However, tests are only one component of vocational guidance interventions. Additionally, Viteles (1936) criticized the failure to include an untested control group and several other aspects of the criteria. E. L. Thorndike (1935) responded to the criticisms. He defended the coarse grouping of occupations and choice of criteria, stated what type of advice to 14-year-old youths could be scientifically supported, criticized the Birmingham study cited as evidence by Macrae, and argued for the use of empirical facts to determine vocational guidance efforts.

With few exceptions, empirical investigations measured only one or two criteria, despite the realization of many prominent writers that the criterion space was, at least conceptually, multidimensional (cf. Peel & Alexander, 1936). There was some attention to reliability of the criterion measure, although Anderson (1929–1930) found that, of 14 clerical tests, four developers failed to report any information on the criterion measure used. Reliance on a fallible and deficient criterion has tended to characterize validation research since its inception (Jones, 1950; Lent et al., 1971). Studies that did measure multiple criteria (e.g., Bird, 1931) were then faced with weighting or combining measures to form a composite (Burtt, 1926), with the available methods being standardized, unit-weighted, or a regression-weighted combination. Another tactic was to use several criterion measures and to validate predictors separately against different measures, as illustrated by Pond (1927) and Wechsler (1926–1927).

As an example of the use of multiple measures in one study, Otis (1938) reported a study of power sewing machine operators (in training) that used multiple work samples as criteria. The work samples, based on a careful job analysis, consisted of the following hierarchical sequence of tasks: straight stitching (paper), straight stitching (material), patch sewing, bias strip attachment, and dressmaking. Three samples were objectively scored, and two were rated by experts. Reliability coefficients for the three tasks that were objectively scored were estimated from the results of two trials of each task and then boosted with the Spearman-Brown formula (the result was .84); they were estimated with interrater correlations for the remaining two tasks (three and four raters and resulting averages of .80 and .75, respectively). Time-to-completion scores were available for four of the tasks, but the observed correlation of −.17 led Otis to analyze quality and speed criteria separately. The work samples were formed into composites on the basis of standardized scores and were related to an array of general and specific ability scores, with resultant batteries identified for quality and speed. This study is exemplary in reporting a large amount of detail concerning the psychometric attributes of the criteria and in using a composite of objective and rated work samples. A similar study by Kornhauser (1923–1924) examined the output of billing clerks, using criteria of speed, accuracy, and ratings of speed and accuracy. Bird (1931) combined non-output measures, namely, salary, number of months employed, salary increase, supervisory ratings, and number of promotions, into an efficiency index.

Mace's (1935) series of studies for the Industrial Fatigue Research Board involving financial incentives represented an attempt to include motivation as a latent cause of observed performance differences. His separation of performance determinants into capacity and will-to-work classes represented one attempt to move beyond the single-facet conceptualizations (on the predictor side) that dominated the thinking of this time.

Methods

Hull (1928) listed product, action, and subjective impression methods for observing work performance, as well as several methods of scoring (counting, objective and subjective scales, ranking, and qualitative scales). The former two methods, product and action, correspond to the well-known results-of-behavior and behavior types of measures, respectively, whereas the latter represents the traditional rating scale method. The following discussion is segmented into soft and hard criteria sections, in accordance with the distinction made by Smith (1976).

Soft criteria. Rating scales were adapted for measuring job performance during this period. We should note Freyd's (1923) acknowledgement of several precursors, including scales pro-

posed by Galton in the 1880s to appraise mental imagery, by Downey for assessing willpower through resistance to opposition, and by Plant for measuring the attention of psychiatric patients. In addition, J. D. Hackett (1928–1929) described ratings of legislators on a "scale of parliamentary merits" published by the Dublin Evening Post in 1784 that closely resembled early graphic scales. Hull (1928) discussed several rating scales under the subjective impression category. Rating scales represented a major technical innovation of those times for assessing performance, particularly the graphic rating scale (e.g., Freyd, 1923; Kingsbury, 1922; Kornhauser, 1926-1927; Paterson, 1923a, 1923b). Its history is as follows. Before World War I, Walter Dill Scott had pioneered a one-to-one scale for the evaluation of salespersons, in which rated individuals were compared with standard salespersons representing five levels of job success. During the war, Scott directed the Committee on the Classification of Personnel in the Army and adapted that format to rate Army officers on five traits: physical qualities, intelligence, leadership, personal qualities, and general value to the service. Each ratee was referenced to a standard scale of five officers derived by each rater (Symonds, 1931, pp. 60–61), with values assigned to each officer-anchor to provide a score. Although quantitative comparisons could have been made to check on agreement in anchor selections or stability over time, we found few evaluations in the literature. However, Strong's (1955) tribute to Scott gave a World War I anecdote about the rating of all colonels and generals (up to but not including Pershing) by a panel of 10 general officers. In an unofficial validation, Strong noted that 9 of the 10 highest rated generals had received special honors or promotions, whereas all 10 of the lowest rated officers had been sent home. Rugg (1921, 1922) analyzed the data resulting from the one-to-one scale during World War I and concluded that ratings were not likely to be useful because of statistical and practical problems (e.g., halo). However, he recommended that at least three independent raters be used, that ratings should be made in conference and under supervision, and that raters should be competent to rate (or have opportunities to observe).

The one-to-one format proved too unwieldy for practical use (Paterson, 1923a; Scott & Clothier, 1923; see Ross, 1966, for an application) and thus served as a stimulus for the development of the graphic rating scale (GRS) during and after the war. The GRS for rating job performance was developed by Ruml (according to Paterson), working with Freyd and Paterson, for the Scott Corporation. The GRS was a numerical scale, with verbally defined anchors, that could assess overall and specific dimensions of performance. A similar scale for the assessment of abilities had been evaluated by J. B. Miner (1917), who reported correlations between single judges as high as .65, between pairs of judges as high as .79, and between traits as high as .85.

In general, empirical research on ratings within organizations (Kornhauser, 1926–1927; Richardson & Kuder, 1933) paralleled research on ratings in other areas of psychology (e.g., Conrad, 1933; Marsh & Perrin, 1924–1925; Newcomb, 1931; Shen, 1925, 1926, Symonds, 1925). There was a fair amount of cross-citation, indicating awareness by I-O researchers of developmental (Weiss, 1933), social-personality (Symonds, 1931), *and educational (Shen, 1925)* domains. At the same time there was recognition of several problems of rating methods (Knight & Franzen, 1922; Rugg, 1921, 1922; Symonds, 1924, 1925, 1931; E. L. Thorndike, 1920). Such problems comprise the negative side of a trade-off involved in using human observers. E. L. Thorndike (1920) extensively discussed the "halo error" that had been identified by Wells (1907) in his study of literary merit of American authors. An empirical extension was provided by Knight and Franzen (1922), who discussed leniency and halo errors within the framework of ratings collected under three instructional sets: self, ordinary, and ideal. The resulting scores were used to create multiple indices for examination of leniency and halo biases.

Kingsbury (1922) studied graphic ratings of 450 bank employees provided by 45 managers and concluded that the ratings suffered from halo, leniency ("high markers") and severity ("low markers"), central tendency, and shifting standards. In contrast to those who emphasized rating scale format as a determinant of the quality of ratings, he contended that equal attention needed to be paid to rating behavior. He proposed an early form of frame-of-reference training involving multiple practice ratings, comparison of ratings to other "incidental criteria" (e.g., records of errors or accidents), and discussions among raters until the ratings were deemed sufficiently reliable for whatever purpose they might serve (p. 383). Kingsbury believed that no quick fix to rater errors was forthcoming and that training would require at least several days and perhaps months to achieve reliable ratings, but should be required (see Kingsbury, 1925–1926).

Symonds (1924) studied the reduction in reliability associated with decreasing number of scale points, recommending that seven categories were optimal. Tracing this idea through history, Symonds's advice was disputed by Champney and Marshall (1939) and subsequently stimulated researchers during the 1950s and 1960s (Bendig, 1954; Komorita & Graham, 1965), whose results in turn led to Monte Carlo investigations (Cicchetti, Showalter, & Tyrer, 1985; Lissitz & Green, 1975). Symonds provided a comprehensive summary of problems and solutions associated with rating methods, discussing chance and constant errors as respective threats to rating reliability and validity (Symonds, 1931). Finally, Vernon's (1938) comprehensive report for the Industrial Health Research Board reviewed the measurement of attitudes (group and individual levels), traits, word associations, and interests. He provided eight conclusions and recommended increased use of factor analysis to study the structure of rating responses.

A three-part series of studies examining graphic rating scales was conducted at the University of Chicago by Kornhauser (1926–1927). Three groups of undergraduates were rated by varying numbers of instructors using 5-point scales for seven traits. Average ratings of seniors, presumably better known by the rater-instructors, intercorrelated on average higher than those of the other two groups who were less well known to the instructors (presumably because of greater variability). In addition, a series of analyses by Kornhauser led him to conclude that more than 4 raters were unnecessary for purposes of achieving adequate rating reliability (similar to Rugg's analyses of the army rating scales and subsequent recommendations). The second phase examined rater agreement, using mean and variability as well as correlational statistics, on the same set of traits.

This phase indicated that interrater agreement was lower (.41 and .38) than intrarater agreement (.60).

The final part of this study examined relationships among the seven traits, using univariate distributions, correlations between the traits, and correlations between averaged rated traits and averaged grade-point average (GPA). Those results indicated that some traits were rated more reliably than others, that the average intercorrelation was approximately .60 for the set of traits, and that the traits on average correlated .81 with the GPA criterion. Kornhauser (1926–1927) concluded that ratings, although often lacking in discriminant validity, could be improved and gave several suggestions on how to proceed. He also emphasized that the ratings he studied were not collected under the conditions suggested by Rugg (1922) and others, which was deliberately done to give a realistic picture of their psychometric attributes.

Bradshaw (1930) described the development of the American Council on Education (ACE) GRS, comprising seven traits. Students and fraternity members at the University of North Carolina were used for evaluating the scale. The three contributions of the ACE scale were claimed to be a shift to verb anchors, inclusion of the behaviorgram (narrative essay) to provide data on personality, and the choice of a sample of traits to be rated. Satisfactory evidence of reliability and validity was reported, but few implications for usage. Bradshaw (1930) also provided 39 principles divided into trait (6), rater (8), rating (8), and scale construction (17) topics.

Sources for obtaining ratings were predominantly supervisors, as opposed to peers and individuals themselves. Such sources as peers and self-ratings were studied by researchers in related domains (cf. Knight & Franzen, 1922; Shen, 1925, 1926; Yoakum & Manson, 1926) but were not extensively used. Sociometry originated with Jacob Moreno in the early 1930s, and became the basis for the peer nomination technique used during World War II (Hollander, 1954; Kane & Lawler, 1978).

During World War I, Scott's Committee on Classification of Personnel pioneered methods to check claims of skill made by recruits. The purpose of trade tests was to discriminate between novice, apprentice, journeyman, and expert levels and compare the resulting score with the claim of the recruit. Chapman (1921), Chapman and Toops (1919), Robinson (1919), and Toops (1921) described several variants of trade tests, including multiple choice, oral, and performance/behavioral. The performance trade test was a precursor of the work sample and was used more as a predictor. However, Otis (1938) showed that it could also serve as a criterion measure. Trade tests thus served as the precursors of specific ability assessments (i.e., job-knowledge tests) and performance testing. Practically speaking, however, trade tests suffered from the expense required to develop and administer them (Adkins, 1947).

Hard criteria. Objective criteria (product measures for Hull) were left to the organization to determine, usually through the industrial engineering department (thus the continuing influence of scientific management). Because of piece-rate payment schemes, those measures tended to be made carefully by the organization, through daily tallies of productivity. Most writers preferred this category of measure but realized some of its defects. Burtt (1926), as noted earlier, called attention to contaminating factors that could interfere with such measures (cf. Bingham, 1926; Viteles, 1925–1926b). However, there was little study of such criteria, perhaps because their objectivity made them face valid to researchers, managers, and workers alike.

Accidents were another early criterion measure that received substantial attention. This occurred for two reasons. One was because transportation was an important emerging occupation as the infrastructure of the United States expanded during this period using various inter- and intracity rail and automobile systems. Thus business and industry had a stake in safe work places. Relatedly, the safety movement was begun to safeguard drivers and pedestrians. Accidents were especially important for companies providing transportation services. Viteles (1925–1926a) and Shellow (1925–1926) described research at the Milwaukee Electric Railway and Light Company designed to select drivers who would have few accidents. Viteles reviewed a wide range of research, including American and European attempts; Shellow described specific techniques derived from Viteles's review. Snow (1926) reported the use of tests and work samples for taxicab drivers validated against accidents, and Wechsler (1926–1927) reported similar predictors validated separately against accidents and other measures (i.e., earnings, efficiency of earnings, and supervisory ratings). Bingham (1928, 1931) conducted a number of early studies under the auspices of the Personnel Research Federation to study accident rates, as did various investigators connected with the Industrial Fatigue Research Board in England (Greenwood & Woods, 1919; Newbold, 1926). A conclusion reached quickly was that accidents were not normally distributed (Lawshe, 1948) but instead were distributed according to a Poisson form (Cobb, 1940). The concept of accident proneness as an individual difference, proposed around 1920, was influential but controversial (Sheehy & Chapman, 1987).

Categorizing Frameworks

Classifications of criterion measures, or ways to arrange them, were crude. They consisted almost wholly of grouping into objective and subjective classes (cf. Burtt, 1926; Viteles, 1932). The two categories were defined on the basis of amount of judgment required and represented a single facet with two levels representing an underlying continuum. An attempt to classify rating scales was made by Weiss (1933) in a review of rating scales for the study of development. Her system consisted of ranking scales and scoring scales; the latter type was subdivided into simple and weighted scoring systems. At the same time, however, Bingham (1926) pointed out the existence of multiple perspectives from which to view performance, specifically identifying the vantage points of management, supervisors, and workers. Other writers agreed (Stott, 1936, 1939; Viteles, 1925–1926b), but there was no expansion of criterion taxonomies toward multifaceted arrangements.

The evaluation of criterion measures received less attention during the period. Three exceptions should be noted. Freyd (1923) argued that ratings, as criteria, could not be validated. However, he proposed several nonstatistical and statistical standards. The second exception was the work of Farmer (1933), an investigator with the Industrial Fatigue Research Board. His scheme for classification of criteria involved objective measures, judgments of performance, and judgments of

ability. The latter two categories were differentiated on the basis of whether judgments were made of objective performance or of the person (i.e., traits). Farmer proposed that the validity of the criterion measure was important and was too often ignored by researchers, although his concept of validity seemed identical to reliability. Several modern procedures were described, including correlations between objective measures and judgments of ability (.35 on 96 cases) and between judgments of ability and of performance (correlation of .92 on 16 cases). This paper appears to have been the first to call attention to the validation of criteria. A third was a paper by Dickinson (1937), who discussed validity for tests and for criteria. A major argument was that validation could proceed in both directions, implying that predictors and criteria as traditionally conceived are measures in a temporal sequence that extend throughout one's tenure with an organization. Not much attention was paid to evaluation of criteria, however, other than the provision of shopping lists of warnings pertaining to objective measures (e.g., Burtt, 1926) and ratings (e.g., Freyd, 1923).

If we consider taxonomic methods to be important in the description and development of relationships among concepts (Fleishman, 1975; Landy & Vasey, 1984), then those aforementioned early efforts lacked the complexity associated with the "worker in his work" concept (Scott & Clothier, 1923). Vocational psychologists provided diversity by including job satisfaction and withdrawal as criteria for assessing vocational adaptation.

Summary

Unsystematic and largely subjective practices had been characteristic of personnel management during the late 1800s and the earliest part of the century, often under the control of individual supervisors (Nestor, 1986). The increasing size of organizations and management's opposition to organized labor led to the formation of the employment management field (Ferguson, 1962–1965). The people described in this article, working in the area of selection, pioneered performance measurement as important for the scientific study and management of industrial personnel. Its use was mostly for test evaluation, although other purposes were known.

Several forces converged during this period to encourage the organized study of work behavior. Clearly, a direct catalyst for the application of psychology to the study of work behavior was the external influence of World War I. Even though the war was over quickly for the United States, those so-called successes led to considerable enthusiasm for testing, although subsequent decades saw a decline in the United States' military capability and corresponding interest in psychology. However, military decline was more than matched by increased application throughout the private sector. The importance of the war for applied psychology was noted by many contributors to the *Annals of the American Academy of Political and Social Science* (Crennan & Kingsbury, 1923).

A second and less direct catalyst for the study of criteria was functionalism, which stressed the importance of individual differences in behavior and their consequences. The functionalist movement contrasted with structuralist emphases on consciousness by moving from a focus on mental structure to a focus on adaptive behavior (Buxton, 1985; Heidbreder, 1933). Psychological work of this period was strongly influenced by the pragmatic philosophy of William James, Charles S. Peirce, and John Dewey. The implications for functionalism were several. First, scientific pragmatism held that theories and concepts should be evaluated by their effects. Ideas were valid if they worked, that is, the value of ideas rested in their utilitarian consequences. This required attention to the measurement of effects. Functionalism, and also behaviorism, converged on the idea that the study of consequences occupied a central position in the conduct of science. However, the pragmatic utility criterion applies an "if it works, use it" logic that emphasized prediction over explanation and understanding. Cattell (1964) differentiated validity and utility, arguing that the former served scientific and the latter practical aims.

Consistent with the agenda of addressing practical problems through psychology, government and private industry began to apply intelligence testing after the war (Bingham & Davis, 1924; Fryer, 1922, 1934–1935). Not all were in favor of this expansion of testing and classification. Walter Lippmann, a journalist, was a prominent and persistent critic of intelligence testers and engaged in several debates with Lewis M. Terman in the popular press (described by Haney, 1981, and Hothersall, 1990). Moreover, Samelson (1977) concluded that the wartime testing movement had not been as effective as was believed. Intelligence tests often had fairly low validities for predicting job success (e.g., Bingham & Davis, 1924). However, Fryer (1934–1935) attributed their decline in use during the 1930s to "economic causes rather than to a 'disillusionment' of industrialists regarding their value" (p. 323).

Sophisticated psychometrics (factor analysis) seemed reserved for the analysis of predictors whereas criteria were chosen for convenience (Jenkins, 1946). Given this state of affairs it was inevitable for the criterion to be faulted for any failure to predict performance, when perhaps the blame should have been more equally shared. During this period, Bingham (1926) and Viteles (1932) extended the concepts of criteria by arguing that the standards used by employers to evaluate work performance might differ from those used by employees to evaluate their own personal success (cf. Lewin, 1936). In modern terms, the goals or values of the groups differed. This argument implied that there could be at least two sets of criterion measures for researchers to consider when validating predictor measures. Additionally, Walter Dill Scott (1917) had differentiated vocational selection from guidance by the problems involved in validating guidance. Treatments of the concept of vocational adjustment (or adaptation) began to appear (e.g., Kitson, 1925; Shaffer, 1936), with job satisfaction as a criterion. Support for this idea was provided by Stott (1936, 1939), who noted that the National Institute of Industrial Psychology (NIIP) in Britain had maintained such dual perspectives on vocational success since 1921. She cited the use of number of jobs, tenure, reasons for leaving, reports from employers, and satisfaction in one study, but noted problems with each and recognized that some combination of success and happiness would equal vocational adjustment. Several early studies of vocational guidance were conducted by the NIIP (e.g., Burt, Gaw, Ramsey, Smith, & Spielman, 1926; Macrae, 1931). Even now few theoretical models of work take both individual and organizational perspectives into

account (one exception is the Minnesota theory of work adjustment proposed by Lloyd Lofquist, René Dawis, George England, and others; Dawis & Lofquist, 1984, provide the most recent statement of the theory). One implication of considering adjustment or adaptation is a focus on both process and outcome, another is the integration or at least cooperation of vocational and I-O psychologists, and a third is the consideration of an enlarged conceptual domain for criteria. Such a domain includes affective as well as withdrawal outcomes of the work role, rejuvenated today in calls for a general behavioral family (cf. Hulin, 1991).

Early criterion measurement resulted from several influences on applied psychology. Obviously, the practical impetus of World War I and the felt need of American business for better selection devices were important external influences or "pulls." Additionally, the convergence of psychological currents, or "pushes," supported the development of industrial psychology with a focus on criteria for vocational selection and guidance. The GRS exemplified the technical side of criterion measurement and generated multiple reviews of development issues. However, early efforts harnessed criteria to validation through the prediction of traits and outcomes, rather than focusing on work behavior per se and attempting to understand how and why those consequences occurred. This premise would direct and plague criterion research over the next 50 years.

1940–1959

A major reason that so much work on criteria was done during this period was because of World War II and the search for substitutes for combat performance criteria, aided by a mobilization of psychologists for the war effort (Marquis, 1944). Applied psychology proved itself to an even greater extent during this crisis than it had in World War I (Bray, 1948; Flanagan, 1947; Meier, 1943; Office of Strategic Services, 1948; Stuit, 1947). Goodenough (1949) explained the difference in contribution between the two wars as a difference between pioneering techniques and better-established ones. This trial by fire had the effect of popularizing industrial applications after the war, much in the same manner as had happened after World War I.

However, there were also contributions during this interval that were less war-related. Among the prominent nonmilitary contributors were the staffs of the United States Employment Service's (USES) Occupational Research Program, the Social Science Research Council, and the Civil Service Commission's Test Development Section. The USES group, in addition to developing the *Dictionary of Occupational Titles* (*DOT*) in 1939, produced an occupational handbook (Stead, Shartle, & Associates, 1940). Included was J. H. Cooper's (1940) chapter on rating forms, which emphasized that major problems had not been solved and reviewed purposes and formats for department store salespeople, employment service personnel, and cafeteria workers (with psychometric information and discussion of rater training). Otis's (1940) chapter described objective measures, including quantity and quality of production, work samples, tenure, and training. He suggested careful study of the distributions of production measures and also listed several influences on those measures that would be described as contaminants: experience, age-sex effects, education, and working conditions. Both chapters gave examples of the appropriate use of measures. Horst and several associates (1941), working under the auspices of the Social Science Research Council, presented a comprehensive survey of the prediction of personal adjustment within vocational, marital, educational, and criminal domains of behavior. They proposed using advanced statistical techniques, argued for greater attention to criteria across behavioral domains, and suggested the analysis of activities into components and the detailed study of ratings. Adkins' (1947) manual detailed the systematic test procedures used by the Civil Service Commission to develop written and performance tests of achievement. Included was a thorough discussion of the reliability and validity of criterion measures and the preparation and standardization of work samples as predictors or criteria.

Dimensions

Where earlier researchers (e.g., Walter V. Bingham, Harold E. Burtt, Morris Viteles) had developed rational lists of measures and essentially assumed that they represented the dimensions of job success, researchers now began to assemble empirical evidence that complemented rational arguments for multidimensionality. Factor-analytic procedures were applied to rating scale dimensions, first by Ewart, Seashore, and Tiffin (1941), and subsequently by Bolanovich (1946), Grant (1955), and Creager and Harding (1958), among others. Those studies invariably indicated multidimensionality.

In 1946, Rothe began an extensive study of output rates among groups of production workers as diverse as butter wrappers, chocolate dippers, welders, and coil winders (specific studies are cited in Rothe, 1978). The studies continued until 1970, with additional research years after the first study (Rothe, 1978; cf. Vinchur, Schippmann, Smalley, & Rothe, 1991). This programmatic work led Rothe (1978) to conclude that incentive systems serve to reduce variability in performance.

The study by Ewart et al. (1941), using 1,120 factory workers, analyzed 12 rating dimensions using L. L. Thurstone's (1931) centroid method. Three factors were interpreted. The first factor was named *Ability to Do Present Job* and had high positive loadings on all dimensions; the second had low positive loadings on safety, job knowledge, versatility, and accuracy dimensions and was tentatively named *Ability Over and Above Present Job;* the third had a large loading on health, but was discounted by the authors because of the unreliability of this rating dimension.

Rush (1953) analyzed 13 criteria measured on 100 salesmen, including three objective measures, nine ratings measures, and one training measure. The objective measures were corrected for opportunity bias by standardizing within four groups of branch offices with similar achievement percentages of quota. Four factors accounted for the pattern of intercorrelations. He named those factors Objective Achievement, Learning Aptitude, General Reputation, and Sales Technique/Achievement. The factor loadings identified subcriteria, used to form four composites that were regressed on 14 predictor measures. The first three composites resulted in significant R^2 values, and the weaker fourth composite was not significantly predicted. Factor analysis was recommended, as was cross-validation of the

results, although as with many factor analysis studies of this time there did not ensue a programmatic series of investigations.

On the rational side, there were continued listings of measures and their specific problems. Toops's (1944) simultaneously acknowledged the practical necessity of a unitary measure of occupational or educational success while arguing that success was not unitary (as illustrated by his example of Typists A and B, one of whom was quick but inaccurate, the other slow but accurate). Toops (1959) contended that it was important to predict profiles of success rather than a composite. Three problems included (a) variables in the profile, (b) units of comparability, and (c) weighting of the components. The advance represented in this review was the realization that success was multidimensional and that a profile of criterion measures should be the goal of prediction. The necessary statistical tools were just becoming available to apply this idea in practice. Toops's (1944) essay was also significant in that it provided early evidence of the conflict between the need for a unitary success score for test validation and the observation that "even in simple jobs success is multidimensional" (p. 274). The following statement captures both his discomfort with recommending a unitary success score and his recognition of the dilemma faced by those in practical prediction situations:

> The success, then, of an individual, about which we prate so glibly, is a complex thing, and if it is to be made, artificially, into a unitary variable must be compounded of the weighted sum of the several component parts, as one, simplest, conception of the matter. (Toops, 1944, p. 275)

After noting that choice of variables will depend on the domain of interest, Toops (1944) reviewed wages, production, quality, rate of learning, supervisor judgments, job knowledge, tenure, supervisory ability, and six incidental factors as subcriterion success variables. A potentially useful decomposition of errors into costs of production, detection, repair/rework, and re-production was presented and promptly forgotten. He was well aware of the work of Viteles and Bingham, of the practical problems involved in gathering criterion data, and of contaminating factors that hindered score comparability. A final section detailed several methods for combining scores.

Nagle (1953), like Toops (1944), emphasized the role of rational judgment in determining criterion relevance. His article addressed several additional issues including criterion dimensionality, methods of development and weighting of subcriteria to form a composite, and a discussion of how the term criterion had been used by other researchers. He was skeptical of Bechtoldt's (1947) suggestion that factor analysis might be useful for deriving dimensions of criteria. He did agree with Bechtoldt, however, on the need for repeated systematic monitoring of the criterion and the subcriteria that comprise it. Although aware that criteria were multidimensional, Nagle's discussion of ways to derive a composite appears less forced than that of Toops (1944). Finally, Nagle (1953) argued for the primacy of criteria over predictors, noting that predictors derived their significance from the criteria used. Several articles written during this era sounded a similar refrain and attributed the new interest in the validity of criteria to the experience of military psychologists in World War II (Jenkins, 1946; Patterson, 1946; Van Dusen, 1947; Wherry, 1957). Nonetheless, some psychometricians still resisted the idea of validating criteria unless they also served as predictors (e.g., Anastasi, 1950).

Rational-economic conceptualizations of criteria reflected an alternative approach to representing criteria. Wherry (1950) described six ways that employees could contribute to profits: output, quality, lost time, turnover, training time/promotability, and satisfaction. A more explicit econometric treatment of criteria was detailed by Brogden and Taylor (1950a) who applied cost accounting, which they termed *tracing out,* to develop criterion composites on a dollar scale. This approach attempted to relate all job success measures to a financial metric and foreshadowed Cronbach and Gleser's (1965) subsequent treatment of decision utility. Brogden and Taylor listed 10 major criterion problems and showed how their monetary approach could provide solutions (several of them acknowledged as partial) for the majority of them. In their terms, a large portion of the criterion problem pertained to finding weights for component variables, with contamination, deficiency, and relevance defined in terms of the weights (sign and magnitude) of individual components.

Methods

In addition to a discussion of relevance (validity), reliability, and variable combination, Nagle (1953) defined criterion development as a sequence of defining the activity, analyzing the activity, defining elements of success, and developing criteria for each element of success. The potential for interplay of inductive and deductive strategies, as well as science and values, can be seen throughout this sequence. Another technical innovation was the proposal that criteria be placed on a common dollar scale (Brogden & Taylor, 1950a). This development foreshadowed subsequent interest in the selection utility movement (e.g., Cronbach & Gleser, 1965; Steffy & Maurer, 1988) and interest in practical concerns. However, researchers often failed to consider the implications of different criterion measures for utility estimates. A final technical advance of this period would be important to later developments in constructing and evaluating criteria. Specifically, the critical incident technique (CIT) method (Flanagan, 1954), proved to be a durable methodological contribution and serves as a cornerstone of behavior-based performance measurement systems. It has also been generalized to other areas of psychology (e.g., preparation of APA ethical guidelines and cases).

Attention to ratings continued to grow. Surveys indicated that about one third of the companies polled in the industry used ratings (e.g., National Industrial Conference Board, 1938; Starr & Greenly, 1938–1939). Zerga (1943) described merit rating systems in place at large manufacturing companies and gave suggestions for their construction (cf. Dooher & Marquis, 1950). World War II research resulted in several advances in methods for measuring criteria. One new format was the forced-choice (FC) technique (an insight of Robert Wherry, Sr., according to Sisson, 1948; Staff, Personnel Research Section, 1946). This format attempted to overcome rater biases by developing a nontransparent approach using desirability and discrimination indices. Research on six different FC scale variants indicated that statements arranged in tetrads of equal favora-

bility yet consisting of two items with high discriminability and two of low discriminability resulted in greater user acceptance, was less prone to leniency, and correlated highly with a supervisory ranking criterion (Berkshire & Highland, 1953). The format was also shown to be more resistant than graphic rating scales to leniency when administered under a "for keeps" condition (Taylor & Wherry, 1951). Although not without its critics (e.g., Travers, 1951), military experience led to positive recommendations and attempts to apply the technique after the war (Zavala, 1965). The major problem was the resistance of raters on the grounds of its nontransparency. This serves as an early example of the influence of constituencies on criterion effectiveness.

There were several postwar developments in rating scales. Dooher and Marquis (1950) edited an extensive treatment of merit rating formats and issues; Mahler (1947) presented an annotated bibliography arranged in five categories (general, administration, specific types of ratings, research, and merit rating experiences). A review by Severin (1952) provided evidence on two questions, the substitutability of various criterion measures for one another and the usefulness of classes of predictors in forecasting job performance. His conclusion about the first question, after tabulating results of studies from Dorcus and Jones's (1950) handbook (426 studies reported between 1906 and 1948), various military sources (Army Air Forces, Personnel Research Section—Adjutant General's Office, and Bureau of Naval Personnel) and contemporary literature, was that the different measures could definitely not be substituted. Severin advised against using training grades in cases for which inferences were desired about job performance (based on a median correlation of .20 between training and job criteria).

Other developments were empirical in nature. L. W. Ferguson (1950), of the Life Insurance Office Management Association (LIOMA), supervised the development of a behavioral checklist, using Thurstone scaling, for clerical employees in the life insurance industry. This work is exemplary in its careful scale development. One hundred job-relevant traits were rated for importance by 320 managers. Sixteen traits, each represented by 100 behavioral statements, were selected for the final scale, which consisted of two parallel forms. Norms were developed with a sample of 17,000 clerical employees. Use of the checklist format meant that a personnel clerk could score the ratings; attention to psychometric details was matched by a concern for practicality.

A series of studies on rating scale formats by Taylor and his associates presaged the disenchantment that would occur some 20 years later with rating scale format. In those studies, Taylor's research group sought to investigate whether different graphic rating scale formats consisting of different combinations of the presence or absence of trait names, trait descriptions, and behavior descriptions would systematically influence the quality (i.e., reliability, leniency, halo, and variability) of ratings. In the first study, R. S. Barrett, Taylor, Parker, and Martens (1958) found that a graphic rating scale consisting of trait titles and behavioral anchors was superior to other formats, including a more structured one that omitted trait titles but incorporated trait definitions and behavioral anchors. In the second study, Taylor, Barrett, Parker, and Martens (1958) grouped the rated traits according to whether they were more person- or job-related (i.e., required less of a moral judgment of the individual and more in the way of observation). Ratings of job traits (e.g., quality of work) demonstrated greater agreement among raters than did ratings of traits (e.g., conscientiousness) for ratings conducted with a less structured scale. In addition, ratings of person traits were significantly more lenient than those for job traits, suggesting that raters were reluctant to rate ratees low on traits that implicated personal qualities. The third study found that self-ratings were significantly more lenient than supervisory ratings (Parker, Taylor, Barrett, & Martens, 1959). According to Parker et al., "These data question the value of using self-ratings as a basis of performance review discussions" (p. 49). Multiple regression analyses indicated that trait self-ratings explained less variance in the overall ratings (69%) than did supervisors' ratings of subordinate performance on the same traits (88%), suggesting that ratees attributed 31% of their performance rating to factors not present on the rating scale.

The final study by Taylor, Parker, and Ford (1959) was perhaps the most significant. This study sought to investigate whether "climate" (e.g., span of control, guaranteed tenure, personnel selection program) differences between four organizations would result in rating differences not explained by scale format. The reliability, leniency, variability, and predictability of ratings varied across organizations and rating formats that were superior by one measure of rating quality were found inferior by other measures. Taylor et al. concluded that situational factors and rater attitude and expectations may better explain rating quality than format differences and they were skeptical that supervisory ratings could be of sufficient quality for validation purposes.

The importance of methodological controls over criterion measurement was another point that received support during the war (Grings, 1952). The classic illustration is found in the evaluation of bombardier performance in terms of circular error, or how close bombs come to a target point (a measure that supports applicability to both practice and real bombing as well as film recording for reliable evaluation). Correlations between bomb runs ranged from −.08 to .37 with a median of +.08 and prompted further investigation (R. Thorndike, 1949). A model of bombardier performance revealed that several situational factors accounted for variability in performance. Factors identified as important included weather conditions, crew performance (pilot, navigator), and equipment variations.

Work samples became more widespread as predictors and criteria, despite well-known time and cost trade-offs (Adkins, 1947; Flanagan, Fiske, Bass, Carter, & Kelly, 1954; Ryans & Frederiksen, 1951). Viteles (1945) summarized research on the "check ride" as a criterion in the evaluation of aircraft or submarine crews. A book describing the Office of Strategic Services (OSS) assessment centers (OSS, 1948), directed by Henry A. Murray, Donald MacKinnon, James G. Miller, Donald Fiske, and Eugenia Hanfmann, illustrated the application of molar principles to behavioral assessment (derived from German and British precursors). Murray and associates summarized the assessment of 5,391 applicants over a 1-day or a 3-day period, with 1,187 of those applicants followed up overseas using supervisor and peer ratings. The use of basic personality principles and careful observations helped enhance the reputation of this method and laid the foundations for Bray and asso-

ciates' implementation at AT&T. However, the attention focused on the predictor side outweighed the attention paid to criterion measures (cf. Holt & Luborsky, 1958; Stern, Stein, & Bloom, 1956). Flanagan (1956) described how the American Institutes for Research had used work samples with the CIT in several organizations.

R. L. Thorndike and Hagen (1959) reported on a longitudinal study of some 17,000 aircrew reassessed more than 10 years after initial testing (1943). Questionnaires, returned by nearly 10,000 of the tested group, assessed present job (to permit DOT coding), income, perceived success/satisfaction, and supervisory responsibilities, as well as work history. Seven criterion measures were derived from the questionnaire, of which monthly income was used in the analyses. Even with the large sample, it was often necessary to group together several DOT codes to secure adequate subsamples. The results, which indicated near-chance validity coefficients for 20 predictor scores derived from the test battery, have not been examined as thoroughly as the quality of the research design might warrant. A major problem appears to have been severe range restriction (mean Army General Classification Test [AGCT] scores of 113), which could be corrected using current formulas. Those results harkened back to an investigation, discussed earlier, conducted by Thorndike's father (E. L. Thorndike et al., 1934), an observation also noted by R. L. Thorndike (1991). A lesson is that problems of criteria strike prominent and ordinary scientists equally, no matter how well conceived the study.

Categorizing Frameworks

There was a great expansion of classification frameworks pertaining to criteria in general and to rating scales in particular after the war. Efforts extended far beyond the objective–subjective dichotomy of the preceding decades. R. L. Thorndike's work with the Army Air Force Aviation Psychology Program and his summaries conveyed practical knowledge outside the military (R. L. Thorndike, 1947, 1949). Thorndike elaborated what is probably the best-known classification scheme for criteria, consisting of ultimate, intermediate, and immediate levels. His analysis of criterion measures rested on the premise that "the ultimate criterion is the complete final goal of a particular type of selection or training" (1949, p. 121). He went on to assert that the ultimate criterion was multiple and complex in every case, an observation that is borne out by examination of both rational listings and the results of previous and subsequent factor-analytic studies. However, by definition, the ultimate criterion, which implies some sort of longitudinal focus over a job or career (e.g., Dennis, 1954, analyzed scientific productivity over 40 to 60 years), is not available to researchers. Thus, practically speaking, criteria of role performance must be drawn from the other categories, intermediate and immediate, which occur closer to the point of selection or training.

Another point made by R. L. Thorndike was that the ultimate criterion must be determined on rational grounds. Then, as researchers and practitioners moved through the intermediate level toward the immediate level, increasing use of empirical-statistical procedures was justified. This point was illustrated by Vallance, Glickman, and Suci (1953) in their outline of an ultimate criterion for the position of Naval officer derived from the U.S. Constitution. Vallance et al. added several facets, including type of role behavior (technical and human relations skill), operational level of analysis (ship, group, and individual), and role phase (training and on-the-job). This may have been the first explicit multifaceted categorization in the history of criterion measurement, although its implications were not drawn out for the applied community. Bass (1952) presented a discussion of ultimate criteria of organizational worth shortly thereafter, arguing that the concept should be expanded to include the perspective of the individual and of society. Material and social classes of criterion measures were distinguished at the organizational level.

Moreover, problems of translating the ultimate criterion concept into intermediate and immediate measures was left unanalyzed, although subsequent work by Astin (1964) and Smith (1976) would address this issue. One conjecture is that some sort of cascading process might translate higher level organizational goals into subunit and individual goals, which would then be translated into action plans by managers and subordinates (Steers & Porter, 1974). The concepts of goals at individual and organizational levels were not well developed at that time (March & Simon, 1958), therefore the tools to conceptualize and examine the hypothesis were unavailable. Even today there is room for investigation and application of such ideas.

Wherry's (1952) model of rating was remarkably current, drawing on psychometric and cognitive research; it decomposed an observed rating into ratee performance, rater observation, rater bias, and error components. Forty-six theorems were generated and many still await evaluation. Other than a few studies by his students (e.g., Bare, 1954; Clarke, 1956; Freeberg, 1955), the theory languished as a technical report. Landy and Farr (1983), who included an appendix written by Wherry, and Wherry and Bartlett (1982), have recently disseminated the theory. Research to test and refine the conceptions is needed.

Evaluation of criteria became a focus. Brogden and Taylor (1950b) ordered criterion bias into environmental/opportunity, group, and predictor score knowledge categories. The importance of their classification was the guidance it provided for researchers interested in controlling for or studying the effects of those three contaminants, although the focus was on control of a nuisance parameter rather than study of it per se. Other classificatory work was an expansion of the earlier focus on the reliability of criteria (cf. Dickinson, 1937; Farmer, 1933). An article by Bellows (1941) is typical; he reviewed three contaminating influences and proposed a list of six "criteria for criteria" grouped into statistical (3), acceptability (2), and practical effect (1) clusters (Wherry, 1950). Freyd (1923) previously used statistical and nonstatistical clusters of standards. Writing on the basis of his Navy work, Bechtoldt (1947) proposed three standards for evaluating criteria: (a) reliability and discriminability, (b) pertinence and comprehensiveness, and (c) comparability. The latter characteristic referred to equivalence across various facets or conditions of measurement (e.g., time, place, or group). R. L. Thorndike's (1949) treatment included four standards: (a) relevance, (b) reliability, (c) freedom from discrimination, and (d) practicality. Relevance would be considered validity, although its common mode of treatment was rational.

Knauft (1947) proposed a three-part breakdown for ratings that used scale development, scale usage, and scoring opera-

tions. The resulting framework was successfully applied to existing formats. Together with a discussion of five scoring methods for obtained ratings, Knauft evaluated the appropriateness of different methods for three common rating situations. Specific and different recommendations were made for each situation.

Summary

Reliance on technology continued during this period, but there was progress in conceptualizing criteria as indicated by the flourishing of categorization efforts. A balance between empirical and rational thrusts remained to be found. One of the most important contributions, in retrospect, was Wherry's (1952) model of the rating process. A second conceptual contribution was R. L. Thorndike's (1949) "ultimate criterion," which stimulated intellectual debate about criteria and served as a statement of a goal that criterion specialists strive to attain. Thus, the ultimate criterion is a construct used by the researcher as a theoretical "tool and goal" (Marx, 1963). A third important development was the framework offered by Brogden and Taylor (1950b) for partitioning sources of variance and covariance among predictors, actual criteria, and the ultimate criterion. The terms *deficiency, contamination,* and *relevance* have by now become essential organizing concepts for students of criteria. Those developments converged on the idea that job performance is multidimensional and subject to various systematic and random fluctuations during measurement.

The practical prediction situation for industrial and other applied measurement specialists continued to outweigh the conceptual work that was required to advance criterion measurement to the next stage, that of developing models of performance and conducting programmatic research. The elements were in place, for example, in that extraneous, biasing influences that invalidated criterion measures were known to researchers (cf. Bellows, 1941; Brogden & Taylor, 1950b). However, those systematic factors were considered as methodological problems to be controlled rather than as topics worthy of investigation.

This period witnessed a greater awareness of the need for balance between technological and conceptual advances. However, the technology developed during this period addressed so well the needs of researchers and practitioners (e.g., forced choice scales, critical incidents, and factor analysis) that they overshadowed the significance of the conceptual advances. Warnings by Jenkins (1946), Wherry (1957), and Wallace and Weitz (1955) appeared as voices in the wilderness.

1960–1979

One of the noteworthy developments of this period was a shift of attention from output measures and personal traits to behavioral measures as criteria. Researchers developed conceptual criteria to capture performance differences and continued to emphasize familiar trait constructs as the latent causes of performance variability (Campbell, Dunnette, Lawler, & Weick, 1970). In addition, the "conventional appraisal" used early in this period was "concerned mainly with personality and character traits. . . So strongly is the emphasis on personality that 'job knowledge' and even 'job performance' may have only a minor place in the over-all rating" (Dale & Smith, 1957, cited in P. Kelly, 1958, p. 60). However, raters were uncomfortable with describing performance in terms of the personal qualities of ratees (McGregor, 1957). Many appraisal systems of the time represented an "eclectic hodgepodge" of personal traits, outcomes, and work behaviors such as absence (P. Kelly, 1958). All too often such scales were introduced on an ad hoc basis or were borrowed from other sources and misapplied to situations in which they were not suitable (Borman & Vallon, 1974; International Labour Office, 1960).

However, there was an increasing awareness that the focus on personal traits and the use of haphazard rating schemes limited research contributions to solving the practical concerns of managing job performance. One response to the widening chasm between practice and research objectives was the proliferation of behaviorally anchored rating scales. J. P. Campbell et al's. (1970) general criterion model went a long way toward illustrating the complex relationships between determinants of performance and performance at different levels of aggregation. Coupled with the necessary technological contribution of behaviorally anchored rating scales (BARS; Smith & Kendall, 1963), this period would witness a growing concentration on behavioral process aspects as opposed to trait and output aspects of job performance (Kavanagh, 1971).

The first edition of the I-O handbook was published in 1976. Dunnette (1976) and Guion (1976) both underscored the need for applying the scientific attitude to applied domains to develop explanations for work behavior and enhance prediction. Reviews by Ronan and Prien (1966) and Schultz and Siegel (1963) assessed progress in criterion measurement and understanding. Schmidt and Kaplan (1971) reviewed the controversy over the use of a composite criterion versus multiple criteria, stressing that the purposes of measurement could differ and thus make one or the other appropriate. Together, those developments suggested that the idea of criteria as measures susceptible to evaluation was beginning to take hold.

Paralleling those scientific developments, Title VII of the Civil Rights Act of 1964 established the goal of equality of employment opportunity and remains a major influence on personnel selection (Dreher & Sackett, 1983; Guion, 1991; Novick, 1981; Wigdor & Garner, 1982). During the 1970s, standards for employment selection based on the Civil Rights Act were promulgated via the Uniform Guidelines on Employee Selection Procedures (1978), although there were differences between the Uniform Guidelines and the *Standards for Educational and Psychological Tests* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1974). Building on the *Brito v. Zia Co.* (1973) decision which held that performance appraisals, when used to make personnel decisions (in that case, layoffs), were subject to the same standards as "tests," researchers began to consider the evaluation of criterion measures (e.g., Kane & Lawler, 1979; Smith, 1976). In this sense, they were focusing research at criterion measures as proposed by Bechtoldt (1951) and Jenkins (1946).

Dimensions

Conceptual work on criterion dimensionality progressed more rapidly than the empirical work necessary to substantiate

the increasingly multivariate conceptions of criteria. Statistical problems operated to discourage the use of factor analysis. Moreover, confirmatory factor analysis was in its infancy (Jöreskog, 1969), and thus was not widely available or understood by potential users. When factor analysis was used to uncover latent dimensions of performance, it was therefore mostly applied in an exploratory and one-shot manner. An exception was the work of Wherry and his students, who followed up factor analyses conducted for the military (Personnel Research Section, Adjutant General's Office) with a hierarchial analysis of objective sales measures for over 900 insurance agents (Roach & Wherry, 1970). The latter analysis suggested one general factor, five secondary or subfactors, and seven tertiary or group factors. Another conceptually driven study was conducted by Inn, Hulin, and Tucker (1972), who studied criteria for performance of an airline reservations clerk job, measured over time, with three-mode factor analysis. The modes were individual, temporal, and criterial; results indicated several interpretable factors within each mode and provided a demonstration of Ghiselli's (1956) scheme for criterion dimensionality.

An elegantly simple correlational study of criterion relationships by Seashore, Indik, and Georgopoulos (1960) illustrates well the difficulty of developing a criterion theory comparable to those applicable to predictor constructs. Seashore et al. studied relationships among five criteria consisting of one subjective and four objective measures on 975 delivery service personnel located in 27 units. The five measured variables consisted of (a) a rating of overall effectiveness (individual and station levels), (b) productivity, (c) chargeable accidents, (d) unexcused absences, and (e) errors. Seashore et al. noted that this subset was chosen for study from a pool of available measures on the basis of their relevance to the firm and estimated or assumed high reliability. Their results, and those of several who preceded their effort (E. L. Kelly, 1957; Rush, 1953), suggested that the magnitude of the relationships among those criteria were small. In the Seashore et al. study, the size and direction of relationships varied among the entire population of employees, among the 27 similar subunits, and within the subunits. The absence of convergent results at both individual and organizational levels of analysis led Seashore et al. to conclude that different job performance variables may in part be determined by different causal variables or differently weighted common causes. They emphatically rejected the concept of unidimensionality for overall job performance.

Klimoski and London (1974) applied factor analysis to study different sources of ratings. They collected self-, peer, and supervisor ratings on 153 registered nurses from four hospitals and analyzed the resulting 60 × 60 (20 criteria for each rater level: 19 specific and 1 overall) multitrait-multirater (MT-MR) matrix using Wherry's (1959) hierarchical method. One advantage of Wherry's technique is that simple structure is maintained at the lower levels, whereas factor covariances are maintained at the highest level. In their own terms, "one general and five subgenerals were extracted" (Klimoski & London, 1974, p. 447). Their general factor (A) was interpreted as indicating that supervisors and peers overlapped in their ratings (cf. Harris & Schaubroeck, 1988). Both group-specific and interrater bias components were identified, but were confounded. The first three subgeneral factors were identified as self-, supervisor, and peer bias, respectively. Factors four and five were labeled Job Competence and Effort. The interpretation of those results was made in terms of Guion's (1965) argument that different dimensions may be used at different levels of the organization (note Borman's approach also published in 1974). An implication is that one should not expect convergence from different sources, and that perhaps opportunities to observe performance and the resulting cognitive units (i.e., person schemas) vary between the three sources. An extension would involve adding sources, specifically subordinates or field review specialists (subject matter experts [SMEs]) and examining those facets in addition to the three studied by Klimoski and London (1974).

A criterion development and evaluation study by Freeberg (1976) provides an illustration of an integrated approach to the development of criterion measures. He was interested in developing measures for youth work-training programs, using three principles of multiple behavioral measures, a temporal continuum for measures, and relevance as a fundamental requirement. Two broad categories of criteria were developed: *program completion,* or proximal, and *postprogram,* or distal. On the basis of factor analyses and prediction of distal from proximal criteria, five measures for the former category and four for the latter category were retained. This study demonstrated how a criterion domain can be mapped out using a combination of rational and empirical procedures, moreover, it demonstrated how criterion measurement can be useful to social policymakers (in this case, those interested in evaluating such youth programs as the Comprehensive Employment and Training Act [CETA]). Orthogonal factoring may have forced independence.

Vroom (1964) and Maier (1955) proposed interactive models of performance determinants comprised of ability and motivation factors. However, increasing awareness that other environmental factors influenced the dependent construct, performance, soon reduced the significance of this two-component model (cf. J. P. Campbell et al., 1970; J. P. Campbell & Pritchard, 1976; Porter & Lawler, 1968). As noted by Blumberg and Pringle (1982), examples of how changes made to environmental variables could influence performance were available early in the history of the field, dating at least to the bricklaying system instituted by Gilbreth (1909).

Methods

Several significant methodological advances during this era were overshadowed by Smith and Kendall's (1963) development of the BARS system. Below, we first review the contribution represented in BARS before discussing other advances.

Behaviorally anchored scales. Smith and Kendall's (1963) BARS system combined the Fels Institute rating format for rating children (Richards & Simons, 1941), Thurstone attitude scaling (Guilford, 1954), and Flanagan's (1954) CIT. The significance of BARS was three-fold, including conceptual, technical, and practical implications.

The BARS approach was significant from several conceptual viewpoints. Unlike other rating formats, the BARS format was designed to use the stereotyping capabilities of raters to promote a common nomenclature among raters with respect to labeling observed incidents of performance. It was also unique

in the sense that it was forward-looking, it was designed to standardize not only the rating process but also the observation process (Jacobs et al., 1980). In essence, the idea was to "foster development of valid stereotypes of effective and ineffective performance prior to observation" (Bernardin & Smith, 1981, p. 459). In a most extraordinary way, the BARS methodology took what had previously been viewed as an obstacle to rating agreement (the human tendency to stereotype and simplify; Wyer & Srull, 1989) and applied it to facilitate rater agreement. However, BARS was not immune to the criticism leveled at other rating formats. Schwab, Heneman, and DeCotiis (1975) argued that empirical evidence for the superiority of BARS was not persuasive enough to offset its costs, while Atkin and Conlon (1978) raised several theoretical problems associated with BARS and concluded that research should be directed at studying the cognitive processes of raters.

Technically, the development process and resulting scales provided a method whereby process aspects of performance (i.e., job behavior as opposed to worker traits or job outcomes) could be measured. This advance ushered in a new era of criterion research that would displace the emphasis on traits and redirect efforts to concentrate on work behavior and emphasizing the role of observation in performance rating. The impact of BARS was so pervasive during this period and so fundamentally redirected appraisal research that Borman (1991) likened BARS to an appraisal philosophy.

In addition to influencing how performance was conceptualized and measured by researchers, the BARS method had several implications for appraisal practice. In the first place, it invited participation by job incumbents in the development of the rating scale through the critical incident and retranslation procedures. The extent of user involvement in scale development was unprecedented. Furthermore, the introduction of behavior-based scales provided an appealing alternative to raters who were uncomfortable with making trait ascriptions for performance (McGregor, 1957). Provision of feedback to employees was a further spinoff anticipated. However, research on user acceptability of BARS has been equivocal (Kingstrom & Bass, 1981) and several debates concerning BARS or behavioral expectation scales versus behavioral observation scales (cf. Bernardin & Kane, 1980; Latham, Fay, & Saari, 1980) have muddied the waters. The finding by Greene, Bernardin, and Abbott (1985) that the formats correlate near 1.00 when corrected for attenuation also suggests that scale development is more crucial than format. Comparisons of cognitive processes involved in completing various rating formats would be one way to integrate the cognitive and format comparison research paradigms, perhaps using process tracing methods and formulating quantitative comparisons between rater groups.

Other rating methods. Format comparisons proliferated during this period (cf. Jacobs, 1986), including studies by Bernardin (1977), Bernardin, Alvares, and Cranny (1976), Bernardin, LaShells, Smith, and Alvares (1976), and Latham and Wexley (1977). No less than five different behavior-based scales were introduced during this period including *mixed standard scales* (Blanz & Ghiselli, 1972), *behavioral expectation scales* (BES; Smith & Kendall, 1963), *behavioral observation scales* (BOS; Latham & Wexley, 1977), *behavioral discrimination scales* (BDS; Kane & Lawler, 1979), and *behavioral standard scales* (Borman, 1986). Though format comparisons predominated, there was no dearth of research on other components of appraisal, notably rater training (Brown, 1968; Latham, Wexley, & Pursell, 1975) and individual differences (Borman, 1979b).

A brief description of a study comparing formats conveys the flavor of this period. Bernardin, Alvares, and Cranny (1976) compared BES and two different summated rating scales (SRS-1 and SRS-2), used to evaluate college instructors, in terms of leniency, discriminability, interrater reliability, and consistency across dimensions. Bernardin et al. noted that previous efforts to compare BES and other formats were limited with respect to the number of psychometric criteria used and that the results across studies were inconsistent (e.g., Borman & Vallon, 1974; J. P. Campbell, Dunnette, Arvey, & Hellervik, 1973; Fogli, Hulin, & Blood, 1971). The Smith and Kendall (1963) procedure was followed for developing the BES, whereas two different procedures were used to develop the summated scales. The first summated scale, SRS-1, consisted of 39 items designed to represent the nine performance dimensions previously identified by a panel of 15 students. A reduced set of 24 items constituted SRS-2 and were based on an analysis of item variances and item–dimension correlations. Bernardin et al. found (a) the BES ratings to be significantly less lenient than the SRS-1 ratings but significantly more lenient than the SRS-2 ratings, (b) that the three scales did not differ with respect to discriminability, (c) that SRS-2 ratings were superior to BES ratings with respect to interrater agreement, and (d) that ratings were more variable across dimensions for the BES versus SRS-1 but did not differ between BES and SRS-2. The findings of this study suggest that rigorous scale development, regardless of scale format, will result in scales that are psychometrically superior to alternatives.

Additionally, a well-designed longitudinal investigation of sources of rating bias was notable for addressing several practical problems of criterion measurement (Campbell, Crooks, Mahoney, & Rock, 1973). Borman (1975, 1977, 1978, 1979a) conducted a program of research on rating accuracy and rater training that stimulated the widespread use of videotape technology for performance appraisal research. In fact, the widely used "Borman tapes" were used throughout the 1980s (e.g., Buckley, Villanova, & Benson, 1989). Bernardin and Walter (1977) evaluated the use of diary-keeping to improve the accuracy of appraisal data. Research on contrast effects in performance ratings (Grey & Kipnis, 1976) and interviews (Carlson, 1967; Rowe, 1967) heralded the recent interest in the effects of the social context on performance ratings (cf. Dipboye, 1985; Villanova & Bernardin, 1989).

Following several studies that found support for training as a means of reducing rating errors (Bernardin, 1978; Borman, 1975; Brown, 1968; Latham et al., 1975), Ivancevich (1979) conducted a study to examine the longitudinal effects of rater training. Engineering supervisors ($N = 66$) were randomly assigned to one of three groups: (a) an intense training group, consisting of a combination of rater error training and frame of reference training with participants viewing 30-min videos depicting below- and above-average performance, (b) a discussion group, during which members reviewed rating errors and discussed sample ratings; and (c) a control group that received no training. When ratings collected 6 months after training were

compared, Ivancevich found that both training groups demonstrated less halo than the control group and that the intense training group's ratings were less lenient than the others. Of more significance, however, was the finding that the beneficial effects of training on halo error dissipated over time, suggesting the need for refresher training, at least for halo error.

Considerable research on individual differences of raters and rating accuracy occurred during the late 1970s. Borman (1979b) reported that the attributes intelligence, freedom from doubt, detail orientation, and investigative interest correlated positively with performance rating accuracy among a sample of undergraduates. The 16 personality variables investigated in this study were chosen on the basis of previous research and informed judgment about traits likely to influence interpersonal perception; thus, those findings suggest that research on rater personality differences may be of practical significance to rater training programs (Heneman, Wexley, & Moore, 1987). Rush, Phillips, and Lord (1981) found that the memory capacity of undergraduates correlated positively with leader rating accuracy. On the other hand, research repeatedly failed to support Schneier's (1977) hypothesis that cognitively complex individuals showed greater rating accuracy (e.g., Bernardin, Cardy, & Carlyle, 1982; Lahey & Saal, 1981; Sauser & Pond, 1981).

Several tactics were used to model judgments of performance. Policy capturing was applied to performance appraisal (Zedeck & Kafry, 1977). In some instances, researchers combined policy-capturing methods with clustering algorithms to group the resulting "paramorphic" representations of judges' policies (Naylor & Wherry, 1965; Wherry & Naylor, 1966). However, although these linear models of human judgment performed accurately in the sense that they explained significant variance in judges' policies, they were faulted for their inability to describe the actual processes that generated the judgments (Anderson & Shanteau, 1977; Slovic & Lichtenstein, 1971), which boils down to an issue of cue validity (Hobson & Gibson, 1983).

Another methodological offshoot involved translations from basic psychology. One such translation consisted of extensions of Thurstonian psychophysics (e.g., Marble, 1942; Richardson & Kuder, 1933; Uhrbrock, 1950, 1961), which define the rater's task as a stimulus-centered judgment, to the evaluation of performance. Whitlock (1963) evaluated Stevens's psychophysical power function for the judgment of effective and ineffective "specimens" of performance across seven samples. He found a good fit of the function, $y = kx^n$, to the judgments of overall performance, with x representing the number of specimens of effective performance, k the ratio of effective to ineffective specimens, and n the number of trials. He called for further research to investigate the effects of rater variables and several aspects of the specimens (criticality, homogeneity, and dimensionality) on the magnitude of error variance in fitting the power function to the ratings. As far as we can tell, no one has extended those ideas, which appear to be related to an article by Garner (1960) that applied information theory to the rating scale. Another translation from basic psychology was the proposal that signal detection theory be used to model the performance rating process (Baker & Schuck, 1975). That proposal did stimulate additional work, including Lord's (1985) merger of schema concepts and signal detection to the issue of rating accuracy, but there has been little empirical work to evaluate the ideas generated. Murphy (1991) recently discussed a related distinction between behavioral accuracy and classification accuracy that may have implications for practice, but evaluation is required.

Smith (1976) reviewed standards for evaluating criteria, relevance, reliability, and practicality; Kane and Lawler (1979) added structural and operational discriminability standards (pertaining to variability among ratees) to reliability and validity (relevance) and further elaborated the three-way analysis of variance (ANOVA) framework for criterion evaluation (Kavanagh, MacKinney, & Wolins, 1971). This approach led to a flurry of multitrait–multimethod (MTMM) studies (e.g., Dickinson & Tice, 1973; Zedeck & Baker, 1972), with methods defined by rating sources (following Lawler, 1967). Holzbach (1978), for example, used a sample of managerial and professional personnel and obtained ratings from self-, superior, and peer sources; he also reanalyzed several published studies. The results of his study, and of the others, usually indicated some convergent validity, almost invariably a lack of discriminant validity, and a strong halo effect. Holzbach attempted to remove the halo effect by using a rating of overall performance and constructing residualized correlation matrices (cf. Landy et al., 1980). Studies that used truly different methods, in order to achieve the method independence recommended by Campbell and Fiske (1959), were rare. Problems with the MTMM approach, described by Kalleberg and Kluegel (1975), can be partially remediated by the use of covariance structure models (Schmitt & Stults, 1986). Another informative study would consist of a quantitative summary across criterion MT-MM studies, as shown by Cote and Buckley (1987) in the study of variance accounted for by trait, method, and error variance across 70 studies. Results would guide research efforts toward larger sources of variance.

Categorizing Frameworks

Developments with respect to categorizing frameworks for criteria were extensive. Weitz (1961) suggested that time, type, and level facets be used to classify and select criterion measures. A comprehensive model of managerial effectiveness was proposed by J. P. Campbell et al. (1970) in which criteria were classified in terms of their proximity to organizational goals. This model, a general paradigm for organizing research along person-process-product lines, was later merged by James (1973) with the multiple criterion model. Smith (1976) proposed a three-faceted classification of criteria along time span, specificity, and match to organizational goals. Dunnette (1963a) criticized reliance on "the" criterion; his multipredictor-criteria-subgroup approach to validation advocated a finer grained analysis of selection systems so that they may be tailored for specific classes of applicants (Dunnette, 1963b). Each of those works converged on the recognition that an ultimate criterion is multiply determined and complex in almost every instance (R. L. Thorndike, 1947; Toops, 1944).

Astin (1964) proposed criterion-centered research as an alternative to construct validity. He elaborated the nature and role of criteria using a distinction between conceptual criterion and criterion performance, applied the distinction to test development, and contrasted his model with construct validation. The

importance of rational judgment of performance measures was stressed, echoing earlier statements by R. L. Thorndike (1949). Astin (1964) insisted that criteria could not be validated empirically (p. 811). On the other hand, Guion (1976) was equally insistent about the need to derive valid and oftentimes multiple criterion constructs in the hypothesis-testing endeavor that he termed validation. His point was that selection research and practice involve developing and understanding of the job and relevant performance dimensions, deriving measures of those dimensions, linking individual characteristics required by the job to specific predictor measures, and using multiple validation strategies to examine the tenability of the working hypotheses.

Ronan and Prien (1966) comprehensively reviewed the literature on criteria, organizing their review around key problems and attempting to inspire a criterion theory. Their four problems were the reliability of performance itself, the reliability of observations of performance, the dimensionality of job performance, and the influence of situational factors on performance. Ronan and Prien maintained that reliability was the key standard for evaluating the adequacy of criterion measures, implicitly denying validation as had Freyd (1923) and Astin (1964). In their conclusion they advanced 15 propositions. Wernimont and Campbell (1968) reframed the concept of behavioral consistency, based on work by Goodenough (1949, p. 93), and argued for the increased use of samples of behavior rather than signs (cf. Schmidt, Hunter, & Outerbridge, 1986; Wigdor & Green, 1986).

Progress with respect to measuring such hard criterion measures as withdrawal and productivity came in the form of reviews by Porter, Steers, and their associates (Porter & Steers, 1973; Steers & Rhodes, 1978). Price (1977) provided a sociological perspective on turnover that added macrolevel determinants. Dalton and Todor (1979) distinguished functional and dysfunctional turnover and Ilgen and Hollenback (1977) demonstrated the advantages of aggregation for representing low base-rate criterion measures. Mobley (1977) proposed a process model that linked employee attitudes to turnover through several intermediate stages, including various cognitive processes and behavioral intentions to search for a new job and to quit. Subsequent conceptual and empirical work by Mobley (Mobley, Griffeth, Hand, & Meglino, 1979) and others (Mowday, Koberg, & McArthur, 1984; Muchinsky & Tuttle, 1979) expanded this model. Bluedorn (1982) integrated the work of Price (1977) and Mobley (1977) with "organizational commitment" and provided a path-analytic test of the model and a cross-validation. His results were reasonably supportive, but suggested some repositioning of variables.

Finally, several developments pertaining to ratings were proposed. Baggaley (1974) classified rating methods according to the "subject-centered approach" and arrived at three clusters: single-trait, single-subject; single-trait, multiple-subject; and multiple-trait, single-subject (cf. Bernardin & Beatty, 1984). Models of rating were proposed, including a cognitively oriented one (Borman, 1978) and a motivational one (DeCotiis & Petit, 1978). Borman's (1978) model comprised three steps—observation, evaluation, and weighting—to arrive at performance dimension ratings. The model of DeCotiis and Petit (1978) posited that ratings were a function of rater ability, rater motivation, and judgmental norms for performance. Research tended to examine different components of those models, in particular concentrating on the second step of Borman's (1978) model and the rater ability component of DeCotiis and Petit's (1978) model.

Summary

Increasing numbers of conceptual analyses, among them those pertaining to the dynamic nature of criteria (Alvares & Hulin, 1972), the use of composite versus multiple criteria (Schmidt & Kaplan, 1971), and the development of multifaceted criterion taxonomies (J. P. Campbell et al., 1970; Smith, 1976) supported greater elaboration and specificity of models of performance and performance appraisal (Borman, 1978; DeCotiis & Petit, 1978; Wherry & Bartlett, 1982). Although multivariate models were in vogue among theoreticians (i.e., Dunnette, 1963b), little multivariate research was conducted to assess the validity of the proposed models (what existed was described by Sells, 1966). The exception, factor analysis, was used in an exploratory sense (cf. Klimoski & London, 1974; Roach & Wherry, 1970). Criteria came to be widely accepted as multidimensional and multiply determined. Empirical efforts to test and refine such multivariate conceptions have recently appeared as the necessary statistical and methodological tools were developed and disseminated (Bobko, 1990; Bollen, 1989; Nesselroade & Cattell, 1988).

Empirical research during this period consisted of building on the advances made during and immediately following World War II. For example, measurement techniques pioneered in the military were extended to civilian contexts (e.g., test batteries, forced-choice ratings, and assessment centers). Research on contrast effects (e.g., Grey & Kipnis, 1976) and illusory halo (Bingham, 1939; W. H. Cooper, 1981) represented the influence of cognitive and social cognitive domains on research and practice issues of I-O psychology.

This period saw rating format research reach its zenith in popularity along with research on rater training and rater individual differences. Feldman (1986) noted that the former two areas represented research in the psychometric perspective, which often neglects the processes occurring during observing and rating tasks. Thus, although rating format studies dominated research on criteria, those efforts failed to identify one rating format that was superior (Greene, Bernardin, & Abbott, 1985; Jacobs et al., 1980). However, Bernardin and Smith (1981) argued that interpretations of studies comparing the effectiveness of different formats require caution because of differences in operations used to develop and elicit ratings among nominally similar formats. The predominance of the BARS method and its judged impact on appraisal research was exemplified in a journal reviewer's comments to a manuscript concerning rating approaches for BARS: "How long are we going to dance on the head of the BARS pin?" (Bernardin & Smith, 1981, p. 458). Landy and Farr (1980) were so disappointed with the format research after wading through it for a review that they called for a moratorium on such efforts. This was based on their conclusion that when rating scales were well-developed and possessed certain critical characteristics (defined performance domain, behavioral anchors, and 3–9 scale points), differences in the

quality of ratings across different formats were negligible. Their contribution was a role-context-vehicle model and orientation toward the cognitive processes of raters.

Throughout this period, it became increasingly apparent to academicians and practitioners alike that adopting any rating format involved a number of compromises with respect to appraisal effectiveness (e.g., Bernardin & Beatty, 1984; Murphy & Cleveland, 1991). For example, although the CIT rating method is effective in providing feedback to individuals, it may not possess psychometric advantages over other formats (Jacobs et al., 1980).

1980–1992

Over the past decade the amount of research on criteria has supported descriptions of a general renaissance in personnel psychology (Smith & Robertson, 1989). As one indicator of this trend, several books devoted to criteria and performance appraisal appeared (e.g., Berk, 1986; Bernardin & Beatty, 1984; Landy & Farr, 1983; Landy, Zedeck, & Cleveland, 1983; Murphy & Cleveland, 1991). Because we believe that recent developments are familiar to many readers, we have chosen to review this period through an admittedly small sample of key topics and issues within our tripartite framework. Topics addressed under dimensionally include rational and empirical investigations, the dynamic criterion controversy, and general models linking predictor and criterion domains. Under methods we examined methods for studying nonnormal data, analysis of covariance structures, meta-analysis, rater training, and several novel approaches. Under the heading of categorizing frameworks, we noted the reevaluations of subjective-objective criteria distinctions and Total Quality Management. Again, we offer general observations before applying the organizing framework.

Several developments in validation occurred after researchers began to grasp the idea that all validity is construct related (Cronbach, 1971; Guion, 1980). Messick (1989) presented the latest thinking on validity and its relationship to values. His central thesis is that validation implies meaning and consequence, or science and values (Messick, 1975, 1980). His key points generalize easily to organizational settings and space is devoted to topics of interest to I-O researchers, especially in the section dealing with criteria. He quotes approvingly Cronbach's (1971) dictum that validation tells about criteria as well as about predictors. It remains to be seen if criterion investigators can or will explicitly incorporate values into their research. A related conceptualization developed by Wiley (1991) can also be related to criterion measurement, even though the focus is nominally on ability constructs and test score variance. Wiley reformulated validity in terms of the abilities and skills required to perform tasks, which are subdivided into life, learning, and test categories. Goal orientation and temporal limitations characterize tasks and permit measurement. In this light, criteria are samples of task performance that come at the problem from another perspective. Wiley's perspective suggests that unified models that incorporate constructs on the predictor and criterion side of the equation will lead to conceptual understanding but do not necessarily address the demands of practitioners and organizations for solutions to measurement problems, at least in the short run. Schmitt (1989) advocated construct conceptions for defining and evaluating criteria, while noting alternatives to the traditional MTMM. Binning and Barrett's (1989) model helped to organize multiple inferences required for validation.

Reviews of linkages between rating tasks and cognitive processes (W. H. Cooper, 1981; Feldman, 1981; Ilgen & Feldman, 1983; Landy & Farr, 1980) marked the beginning of a paradigm shift toward the use of cognitive models and methods to examine ratings. The merger produced several models of the cognitive operations involved in ratings, of which the best-known were proposed by Feldman (1981, 1986) and DeNisi, Cafferty, and Meglino (1984). The general form of these models is sequenced from observation of behavior through encoding, retrieving, and generating the rating (i.e., integration of information). An issue concerns when evaluation of the observed behaviors takes place, at encoding or retrieval stages. DeNisi and Williams (1988) reviewed the various models beginning with Wherry's (1952) and noted that different models tended to emphasize different cognitive stages. Bretz, Milkovich, and Read (1992) recently classified the bulk of cognitively-oriented studies into two groups, those examining the effects of expectancies and of memory, although they did point out the diversity of this research area.

Additional information bearing on this issue was provided by Landy and Rastegary (1989), who used a content analysis of 11 journals over 7 years. They classified criterion articles into five categories (plus a multiple indicator category, which made up about 10% of the total). They were able to demonstrate a limited domain of primary publication outlets for 408 studies that concerned criteria. Landy and Rastegary advanced several reasons for the decline of standard validation studies, which constituted about 7% of the total number of studies over the period of the review; suggested increased publication of research on issues relevant to criteria (e.g., the *Test Validity Yearbook* as a replacement for the *Validity Information Exchange*); and reported conclusions and trends for the five categories of criteria.

The influence of federal legislation on personnel psychology in general (Novick, 1981, 1982), and on performance measurement in particular (Barrett & Kernan, 1987; Cascio & Bernardin, 1981), has increased during this period. Not since the Tenth Circuit Court's decision in *Brito v. Zia Co.* (1973) has there been so much controversy and confusion regarding the defensibility of performance measurement practices in organizations. A series of Supreme Court decisions in the late 1980s (*Price Waterhouse v. Hopkins*, 1989; *Watson v. Ft. Worth Bank and Trust*, 1988), which focused on the use of subjective appraisals for internal selection (promotion), and one that dealt with employment practices (*Wards Cove Packing Company v. Atonio*, 1989), seemed inconsistent in their aims. The Supreme Court held that plaintiffs maintain the burden of demonstrating a causal link between a specific employment practice and statistical imbalance (*Wards Cove Packing Company v. Atonio*); that plaintiffs may challenge subjective procedures under the disparate impact model (*Watson v. Ft. Worth Bank and Trust*); and that once a plaintiff has shown that an unlawful factor played a part in the adverse selection decision involving a mixed-motive disparate treatment case, the evidentiary burden shifts to the employer

(*Price Waterhouse v. Hopkins*). The recently enacted Civil Rights Act of 1991 amended Title VII and established that the burden of proof would remain with the defendant-employer to demonstrate that a selection procedure was "job related and consistent with business necessity." Precise definition of terms awaits further interaction of psychological standards and judicial-legal interpretation (cf. APA, 1989; Lee, 1990).

Dimensions

Dimensionality was again addressed rationally and empirically. Several researchers suggested an expansion of criteria to include extra–role behaviors, defined as those that go beyond the requirements of a job role. In particular, where March and Simon (1958) defined the basic human resource problem for organizations as providing inducements in return for contributions (i.e., participation and production), the new logic proposes that prosocial or "good citizenship" classes of behavior be added (Organ, 1988). Examples of such behaviors might include providing assistance to new members of the organization, as in unofficial mentoring, or working on weekends to complete a project in a timely manner. Several empirical studies provided initial evidence for this expansion, but the focus has been more from an organizational psychological perspective than from a criterion measurement one. Conceptual and measurement issues remain, including dimensionality.

Several factor-analytic studies appeared, including a number based on research conducted for the joint services to validate the Armed Services Vocational Aptitude Battery against job rather than training proficiency (Wigdor & Green, 1986). One component of that effort is the Project A research being conducted for the army (cf. J. P. Campbell, 1990b); another is the Joint-Service Job Performance Measurement (JPM) project (Wigdor & Green, 1986, 1991). In one study, J. P. Campbell et al. (1990) developed a five-factor model of performance that included soldiering and core technical proficiency, physical fitness/military bearing, effort/peer leadership, and personal discipline constructs. They confirmed its structure across several Military Occupational Specialties (MOS), providing some evidence for generalizability, and suggested that portions might prove applicable to civilian job families. Borman, White, Pulakos, and Oppler (1991) posited models to extend Hunter's (1983, 1986) path models. Hunter's models were concerned with the effects of general ability, manifested through specific job knowledge, on work sample scores and supervisory ratings. The extensions tested by Borman et al. consisted of adding technical proficiency and problem behavior measures. Model testing revealed substantial direct effects for the specific job proficiency and problem behavior constructs, as well as indirect effects for ability, job knowledge, and dependability. Comparative evaluation indicated that approximately twice as much variance was accounted for by the expanded model as by the variables advocated by Hunter (1986).

An increasingly important and disputed facet of dimensionality is temporal. The concept of dynamic criteria, proposed by Ghiselli (1956) as one of three facets of criterion dimensionality (the others were static and individual) and empirical work (Alvares & Hulin, 1972; Dunham, 1974; Ghiselli & Haire, 1960), suggested that there were several models capable of accounting for the declines. A series of debates that generated new research and advanced understanding of the phenomenon were initiated by Barrett, Caldwell, and Alexander (1985), who reviewed and critiqued three definitions of criterion dynamism: (a) mean changes, (b) predictor-criterion changes (declining validity), and (c) criterion-criterion changes (declining stabilities). They concluded that dynamic criteria had become what they called a "received doctrine" (Barrett, 1972). Henry and Hulin (1987), using a sample of professional baseball players, found evidence for a simplexlike pattern of declining stabilities and discussed corresponding implications for utility analyses, which assume that validity is stable over some period. Deadrick and Madigan (1990) reported a replication of a longitudinal study by Rambo, Chomiak, and Price (1983), finding support for declining stabilities. Hulin, Henry, and Noon (1990) reported two meta-analyses of the effects of time interval on size of validity coefficient, one for predictive validities and one for criterion stabilities. After making corrections where possible for attenuation and range restriction, their findings indicated strong negative relationships: −.60 for initial-final performance and −.80 for predictive validity studies. These coefficients, which result from regressing validities/stabilities on time, index of the amount of decline proposed by Ghiselli (1956) as one of three facets of criterion dimensionality (the others were static and individual), received increased attention during this interval. Earlier theoretical explanations for the negative relationships were reviewed, including the changing task and changing subject models first elaborated by Alvares and Hulin (1972). Hanges, Schneider, and Niles (1990) applied an interactionist perspective to change in teacher ratings over 6.5 years (13 semesters). Person, situation, and person-in-situation analyses were applied to correlation matrices containing seven specific dimensions and an overall composite, resulting in different patterns over time for specific dimensions but a claim of impressive stability.

Together, these efforts are in keeping with the recommendation of Austin, Humphreys, and Hulin (1989) that empirical research should be driven by conceptual work and modern statistical procedures to provide increased understanding of this phenomenon. Future studies of dynamic criteria should permit the estimation of individual growth parameters as recommended by Rogosa, Brandt, and Zimowski (1982), Rogosa and Willett (1985), and Bryk and Raudenbush (1987). Several recent studies have used this approach. An example is work by Hofmann and associates (Hofmann, Jacobs, & Baratta, in press; Hofmann, Jacobs, & Gerras, 1992), which explicitly rejects the analysis of correlations. Hofmann et al. (1992) studied the performance of two groups of major-league baseball players, hitters (using batting average) and pitchers (using earned-run average), over their first 10 years in the league. Growth curves were estimated for each person using linear, quadratic, and cubic coefficients. Hofmann et al. (in press) generalized this method in a study of a large cohort of life insurance salespeople over 12 quarters. Descriptive growth curves were fit as before, and a hierarchical linear model (HLM) approach was used to investigate "interindividual differences in intraindividual change" (cf. Nesselroade, 1991). Three distinct clusters of individual change patterns were found, and several explanations, one in terms of goal orientation, were advanced. It is also possible to fit latent

growth curve models (Meredith & Tisak, 1990) to such data sets (cf. Browne & DuToit, 1991).

Conceptual models of change were proposed by Murphy (1989b) and Fleishman and Mumford (1989). Murphy's (1989b) model was developed in response to the study by Schmidt et al. (1986), which concluded that general ability is largely responsible for job performance, with both direct and indirect (through job knowledge) effects. Murphy distinguished between transition (e.g., due to acquisition, new technology, or job change) and maintenance stages of task performance, contending that cognitive ability would be most influential during transition stages. The proposed model accounted for the findings of Schmidt et al. (1986) as well as other research findings. The model proposed by Fleishman and Mumford (1989), in response to a critique by Barrett, Caldwell, and Alexander (1989), incorporates a latent predictor construct that influences trial-to-trial increments across different stages of task performance. Multiple indicators permit the estimation of measurement and structural components of the model. Adding motivational constructs to this model might be profitable, as would combining it with the techniques of Hofmann et al. (in press).

A crucial point is the need to move beyond demonstrations of the phenomenon (or nihilistic critiques) to developing and testing explanations. Sources for such theoretical work might include the literatures of educational (Bloom, 1964), developmental (Vondracek, Lerner, & Schulenberg, 1983), experimental (Seashore, 1939), and interactional (Schneider, 1983) psychology. Hull's (1952) concept of behavioral oscillation could serve as another theoretical bridge to the basic psychology literature, as noted by Parker (1971). Work on conceptions of time might also be useful (Katz, 1980; J. Kelly & McGrath, 1988). Additionally, one way to resolve the conflicts between static and dynamic criterion perspectives involves the method used by Latham, Erez, and Locke (1988) to address the dispute between Erez and Latham concerning participation. Specifically, members of both perspectives would be invited to design a study or series of studies to resolve the dispute. Study designs should incorporate large samples working in stable jobs and also should avoid range restriction because of differences in performance (i.e., low performers experience involuntary turnover, whereas high performers are promoted or voluntarily leave). Both objective and subjective criterion measures should be collected. Studies should be designed to illuminate components of systematic change in performance, with a practical implication the observation that formulas for estimating selection utility assume a constant validity over some time period. Cattell's (1988a) conception of a data cube appears useful for conceptualizing facets over which change could occur.

Other developments included the proposal of general models incorporating both predictor and criterion constructs (Campbell, 1990a). Blumberg and Pringle (1982) proposed an interactive model of performance determinants consisting of capacity, willingness, and opportunity terms. The model provides a useful organizing framework that should stimulate discussion of opportunity. However, because so few categories are used to describe the universe of potential causes of performance, there remains considerable conceptual and empirical work. For example, their model does not distinguish between opportunity differences that may affect performance on job functions more central to overall performance versus those that may only be peripheral. Murphy and Kroeker (1988; Murphy, 1989a) presented an approach to work performance dimensionality. Their contention was that complete models of predictor and criterion domains, the latter divided into task and job performance, were required to implement the Navy contribution to the JPM project (Wigdor & Green, 1986). One aspect of their argument was the importance of content validity in developing samples of a larger performance domain. Such an emphasis is consistent with our working definition of criteria, which stresses that criterion measures are samples of a larger performance universe. As an example, Laabs and Baker (1989), using Guion's (1979) procedures for work sample development, demonstrated the development of job sample evaluations for the Navy Radioman enlisted classification using critical tasks. Guion (1979) argued that the sequence includes definition of a job content universe, which is refined into a domain, followed by specification of a test content universe, also refined into a test content domain. Laabs and Baker (1989) used a multimethod strategy that incorporated job analyses, interviews/observations, card sorts, factor analyses, and surveys. Their final list, which comprised communications, setup, maintenance, and security dimensions, was judged as acceptable by a sample of incumbent supervisors and as superior to a set generated by random sampling.

Methods

Advances in the statistical analysis of criterion data have promoted research designed to understand criteria. Hulin (1991) called attention to the problems of studying variables with nonnormal distributions and low base rates, noting that Pearson correlations will not estimate relationships accurately under such conditions (cf. Hulin & Rousseau, 1980; Landy, Vasey, & Smith, 1984). Problems with normal theory models make event history and survival analysis models more appropriate for such data, because they focus on states (here, attendance/absence), time spent in those states, and transitions between states. Fichman (1984, 1988) and Harrison and Hulin (1989) applied such models to the study of absence processes over time. Fichman used an approach based on Naylor, Pritchard, and Ilgen's (1980) comprehensive theory of behavior in organizations. Harrison and Hulin used Cox regression after noting that 25 of 27 articles in four major journals (1982–1986) had used inappropriate normal theory analyses. Their large sample ($N = 2,130$) of white-collar financial services employees, as well as their appropriate analyses, make this a demonstration of a useful method for examining absences. Their results showed that temporal and historical variables predicted the hazard rate, but that demographic variables did not. Hackett, Bycio, and Guion (1989) used an idiographic-longitudinal approach to this topic in a sample of nursing personnel, with univariate and multivariate analyses conducted at individual and at group levels to examine this phenomenon. They found that self-reports concerning reasons for absence and attendance predicted intentions to be absent and actual absences. Those design and analysis techniques, essentially relating parallel time series at the individual level, comprise another method of assessing behavior switching between competing families, attendance and ab-

sence. Extensions to accident or error measures would be straightforward.

A second methodological development was the analysis of covariance structures with measurement and structural models. In an early application, Hunter (1983, 1986) estimated various models relating general ability to specific job knowledge scores, work samples, and supervisory ratings using correlation matrices aggregated separately for civilian and military samples. Subsequently, Vance, MacCallum, Coovert, and Hedge (1988), Vance, Coovert, MacCallum, and Hedge (1989), Schmidt et al. (1986), and J. P. Campbell et al. (1990) reported studies of structural relations among ability and job performance constructs measured with multiple methods and indicators. Recent studies reflect an awareness of Guion's (1983) charge of misspecification because of such missing constructs as motivation and personality. Still, improvements are needed in both conceptualization and analysis of the replacement models, as well as generalization across occupational groups. A start might be made by comparing similar occupations across military branches using data from the JPM project.

A third methodological development, shared with other areas of psychology, is quantitative reviews of research domains (Green & Hall, 1984). The goals of such aggregate analyses are to summarize effect sizes from disparate studies and to make corrections for sampling error, unreliability of measures, and range restriction. Quantitative reviews of criteria have summarized such topics as relationships among different rating sources (Harris & Schaubroeck, 1988), self-estimates of ability (Mabe & West, 1982), the convergence of objective and subjective criterion measures (Heneman, 1986), rater errors and accuracy (Murphy & Balzer, 1989), the use of real versus simulated ratees (Murphy, Herr, Lockhart, & Maguire, 1986), comparisons of different criteria for validating clerical selection devices (Nathan & Alexander, 1988), and the effects on ratings of age (Waldman & Avolio, 1986) and race (Ford, Kraiger, & Schechtman, 1986; Kraiger & Ford, 1985). The full potential of quantitative reviews to advance understanding of criteria requires the provision of guidance for future research, perhaps guided by Slavin's (1986) ideas of "best evidence synthesis."

Rater training received attention during this period that went beyond the how-to sessions of earlier eras. Spool's (1978) review set the stage for the modern era, grouping 27 studies into four areas. Bernardin and Buckley (1981) argued that rater training was no panacea for eliminating rating errors and suggested that the most successful training efforts would be those that provide raters with a common frame of reference with which to observe and evaluate job performance. Bernardin and Pence (1980) surprised criterion researchers and practitioners with the finding that rater error training could result in unintended rater response sets. Pulakos (1984) and Hedge and Kavanagh (1988) showed the superiority of accuracy training over error training. However, few researchers cite the influence of other areas of psychology (e.g., Boice, 1983; Funder, 1987; Taft, 1955) and thus the research could be described as minor variations on the same theme.

Several studies involved novel or previously abandoned rating methods. Komaki, Collins, and Temlock (1987) reported the development and application of an applied operant measure for assessing customer service in a retail clothing department. The method is laudable with respect to the care taken to train observers in order to reach high levels of reliability. However, data bearing on other psychometric characteristics of the method are needed. A second development in performance measurement methods is that of relative rating systems. Miner (1988) described a rated ranking technique administered during the course of interviews with 21 foremen who rated 6 to 19 subordinates in a total ratee sample of 185 skilled and semiskilled workers. His analyses centered on interrater reliability, distributional characteristics, leniency, halo, and construct- and criterion-related evidence. Unfortunately, characteristics unique to this study such as confidentiality of ratings and the use of an interview procedure involving the investigator and supervisors clouded interpretation of results that might otherwise appear favorable.

Kane's (1986) performance distribution assessment (PDA) is an example of a potentially useful method that has implications for all three organizing facets of this review. Briefly, Kane's contention is that cyclic work performance gives rise to a distribution over a given period that permits researchers to assess parameters in addition to level. Two parameters of immediate interest are variability and shape of the performance distribution. All other things being equal, organizations should prefer members who display high levels of performance, low variability, and whose distributions are negatively skewed (avoiding extremely low or negative instances of performance, negative range avoidance in Kane's terms). Another implication of this work lies in the possibility of correcting for situational constraints on performance. A third pertains to implications for the dynamic criterion issue. However, there is much work to be done before the promise of PDA can be realized (Borman, 1991). At a minimum, integrated laboratory and field studies that use longitudinal and individual levels of analysis are prerequisites for further study.

Categorizing Frameworks

Several developments were related to the additional statistical analyses described earlier. In particular, discoveries were made with respect to negative characteristics of such hard criteria as withdrawal and productivity. Hammer and Landau (1981), for instance, showed that empirical research on absence behavior could arrive at different conclusions on the basis of how absence was operationally defined (cf. Hulin & Rousseau, 1980). Kemery, Dunlap, and Bedeian (1989) demonstrated that different definitions of employee separation have statistical implications that can result in different conclusions. Campion (1991) developed measures for five definitions of turnover and found that they differed in degree of independence from each other and relatedness to predictor variables. He concluded that turnover measures conceptualized and measured as varying along a continuum demonstrated higher internal consistency and interrater agreement than did dichotomous turnover measures. This observation is related to a consideration of an underlying "propensity" continuum (Drasgow & Hulin, 1990) and represents an advance over standard binary conceptions of turnover. Peters and Sheridan (1988) pointed out the inferential problems associated with research designs that censor observations of employee flows into and out of the organization. Con-

trary to earlier single-cause explanations of withdrawal, Fichman (1989) suggested that different processes may be responsible for different withdrawal behaviors by identifying positive duration dependence for absence and negative duration dependence for turnover. Hulin's (1991) call for a general behavioral withdrawal construct is relevant here.

Productivity indices were also scrutinized and found to be potentially misleading (Mahoney, 1988), particularly when viewed from systems perspectives (Muckler, 1982). The conceptualization of situational constraints by Peters and O'Connor (1980; Peters, O'Connor, & Eulberg, 1985; Steel & Mento, 1986), and the expansion of that idea into a facilitation-constraint continuum (Schoorman & Schneider, 1988), promoted the incorporation of situational factors within models of work performance (Mitchell, 1983).

Total quality management (TQM; Deming, 1986; Duncan, 1974) has recently captured the interest of American researchers, perhaps because of its antiappraisal philosophy and its storied success in Japan. According to Deming, individual differences in capacity and willingness (adopting Blumberg & Pringle's, 1982, taxonomy) account for trivial portions of performance relative to system/opportunity variability (upwards of 85%). Dobbins, Cardy, and Carson (1991) discussed the implications of adopting a system approach to performance appraisal with liberal reference to TQM concepts. However, primary research on the substantive hypotheses associated with the TQM concept have only recently begun (e.g., Carson, Cardy, Dobbins, & Stewart, 1992). A recent quantitative review by Alliger and Hosoda (1992) involving the analysis of distributions of work output from 101 studies indicated that most tended to approximate normality with the exception of machine operators, a job in which one would expect situational constraints to matter most. This finding, together with research showing that situational constraints act largely as a nuisance variable on performance, suggests that system variance may not be as significant a determinant of performance as TQM advocates make it out to be. On the other hand, Johns's (1991) discussion of substantive and methodological constraints on behavior and attitudes makes the point that reduced criterion variability is a major result of such phenomena. Coordinated lab–field studies will help to illuminate this aspect of criterion variance.

Summary

Widespread concern for the validity of criterion research was stimulated by earlier reviews that concluded that significant boundary variables (Fromkin & Streufert, 1976) might limit the generalizability of laboratory research on performance appraisal to actual appraisal situations (Bernardin & Villanova, 1986; Dipboye, 1985; Ilgen & Favero, 1985; Wendelken & Inn, 1981). Banks and Murphy (1985) warned that the new emphasis on cognitive processes threatened to increase the research–practice gap in performance appraisal. In support, Murphy et al. (1986) showed that effect sizes in appraisal research were significantly larger when the stimuli took the form of "paper people." Thus, prescriptions for appraisal practice based on laboratory research may be of questionable ecological validity for the context in which observation and ratings of work performance actually occur. The findings of Zammuto, London, and Rowland (1982) suggested that organizational differences may act to circumscribe further the generality of the findings of appraisal research across organizational settings; the validity of research findings based on appraisal participants in one applied setting may not generalize to other people and settings. This observation harkens back to Bailey's (1983) contention that criteria are context-sensitive.

Apart from the numerous cautions that boundary variables constrained the generalizability of research on criteria, several researchers argued that the validity of standards used to assess the accuracy and comparability of ratings were inadequate for generating accurate inferences (Bernardin & Smith, 1981; Saal, Downey, & Lahey 1980). Works critical of criterion research comprise a significant development during this period and are in some ways reminiscent of the "crisis of confidence" that the field of social psychology experienced during the 1970s. As the decade closed, a growing number of I-O researchers argued that much of the preceding research on ratings had used inappropriate standards to assess the validity of subjective ratings (Becker & Cardy, 1986; Murphy & Balzer, 1989; Pulakos, Schmitt, & Ostroff, 1986; Sulsky & Balzer, 1988). Conceptual and empirical work on hard criteria did not escape this "criterion crisis" of the 1980s. Rather, much of the criticism directed at subjective ratings was found to apply equally well to popular hard criterion indices such as turnover and absence. Collectively, critical reviews indicate that self-correcting processes of science (Krathwohl, 1985) function among the community of criterion researchers.

Self-correction has occurred also with respect to the direction of criterion research. Subsequent to Landy and Farr's (1980) call for concentrated study of cognitive processes of raters, criterion research moved strongly in that direction. However, the apparent monopoly enjoyed by the cognitive paradigm during the 1980s began to give way as the research–practice gap widened. Thus, by the late 1980s research on criteria began to take stock of cause-and-effect variables related to criterion issues that had since been overlooked due to the preponderance of interest in rating formats during the 1970s and the paradigmatic shift to the study of the cognitive processes of raters during the 1980s. Calls for theories of performance to match those of predictors (J. P. Campbell, 1990a) represent another attempt at self-correction.

On the one hand, there are several cogent critics of the cognitive movement in performance appraisal (Banks & Murphy, 1985; Dipboye, 1985; Ilgen & Favero, 1985), which revolve around external validity and omitted variables. Dipboye (1985), for example, argued that behavioral, social, and motivational variables had been excluded from consideration by cognitive performance appraisal researchers. On the other hand, Landy and Zedeck (1983) pointed out the morass of methodology that had enveloped psychologists interested in ratings for almost 40 years. In sum, then, the jury is still out on the shift toward cognitive models and methods in criterion measurement.

Work on criteria has recently been directed toward understanding of implications that criterion measurement practices have on related HRM subsystems. Some issues representative of a systems view include recent research on appraisal satisfaction (e.g., Dobbins, Cardy, & Platz-Vieno, 1990; Giles & Moss-

holder, 1990; Greenberg, 1986) and format effects on goal setting (Tziner & Kopelman, 1988). As the interest in research on rater cognitive processes reached its apex in the mid-1980s, research attention began to shift toward the study of the difficult issues of rater motivation (e.g., goals) and contextual factors (e.g., sanctions against biased ratings) within a communication framework (Murphy & Cleveland, 1991). This shift is important because previous research tended to focus on the ability portion of the equation predicting rating performance, rather than as a joint function of ability and motivation. Dipboye (1985) and Villanova and Bernardin (1989) presented models of rating behavior that focused on rater motives to distort ratings purposefully. Other topics relevant to rater motivation include the political aspects of the appraisal process (Longenecker, Sims, & Gioia, 1987); perceptions of appraisal fairness (Greenberg, 1986); impression management (Fandt & Ferris, 1990; Villanova & Bernardin, 1989); and the ethics of performance appraisal (Banner & Cooke, 1984). Unlike research on rating formats that predominated during the 1970s and the more recent work on rater cognitive processes, these more recent topics require multiple research strategies. Although the use of actual appraisal participants accords this research higher external validity and facilitates the application of research findings to practice, it also weakens inferences regarding cause-and-effect relations among variables.

Fisher (1989), critical of the slow transfer of knowledge from the laboratory studies of the 1980s to the practice issues of the 1990s, decried the lack of research directed at the needs of human resource management professionals. Dependent variables in cognitive research on performance appraisal consist of various accuracy and error indices, with concentration during the 1980s focused on the measurement of halo and its relationship to rating accuracy. Research on rater leniency, which is arguably of more practical significance for rating utilization, remains primitive relative to research on halo. For example, a survey by Bretz, Milkovich, and Read (1990) found that 77% of respondents from sampled Fortune 100 companies believe that skewed appraisals threaten the validity of their appraisal systems. Lenient ratings also confuse the relationship between true performance differences and reward differences, potentially harming perceptions of pay system fairness (Miceli, Jung, Near, & Greenberger, 1991), one of the most important performance appraisal issues faced by management (Bretz et al., 1992). Over 35 years ago, Glickman (1955) observed that leniency can jeopardize the potential benefits of merit ratings on employee motivation. Finally, leniency has potential legal ramifications as well. For example, courts look specifically at the fairness and feedback provisions of performance appraisals in arriving at verdicts involving wrongful discharge cases (Bernardin & Cascio, 1988). If rating research is to realize its potential for informing practice, then a redirection of efforts toward the study of rating errors and biases that have much more significant operational impact on appraisal systems, such as leniency, seems to be in order. Only recently does it seem that researchers have found it acceptable to trade off internal validity to investigate issues that are of concern to human resources professionals. Cleveland, Murphy, and Williams (1989) provided information relevant to the use of criterion measures for multiple purposes.

Conclusion

Before we proceed with some concluding statements, one caveat is in order. It is all too easy to criticize from the vantage of the present those researchers and practitioners who made initial progress in defining and measuring criteria. Nevertheless, several themes emerging from the review are elaborated on. Suggestions for needed research are given as well.

First, like many subdisciplines of I-O psychology, inductive strategies dominated early efforts to resolve the criterion problem and deductive, model-based strategies played a subordinate role until recently. This is reminiscent of psychology at the time Lewin (1931) analyzed the Aristotelian versus Galilean modes of thought. The former emphasizes classification, the latter explanation. Perhaps this occurs because I-O psychology "has, from its earliest development, placed considerable emphasis on good description and observation to provide the database for the subdiscipline" (Dubin, 1976, p. 18). Given the scant amount of empirical evidence available at the time regarding the nature of the phenomenon (work performance), the pioneers of criterion research were wise enough to pursue an inductive strategy. Assuming the validity of initial premises and observations, the inductive inferences that followed were likely to be probabilistically valid (Salmon, 1967; Salmon, Jeffrey, & Greeno, 1971). However, even though I-O psychology as a scientific discipline has matured to the point where deductive strategies are now commonplace, inductive methods have not yet outlived their usefulness. Locke (1986) and Cook and Campbell (1976) argued that when it comes to assessing the generalizability of research findings, inductive strategies are preferred. Fromkin and Streufert (1976) referred to deductive discussions of external validity as "idle speculation" (p. 457).

Validity of Criterion Measures

The current approach to inquiry about criterion measures, based on validation and embracing a balance of inductive and deductive strategies, can be traced back some 30–35 years to skeptics of the criterion (e.g., Dunnette, 1963a; Ghiselli, 1956), to advocates of a hypothesis-testing approach to validation (e.g., Guion, 1961; Wherry, 1957), and to those who argued for criterion studies "for understanding" (e.g., Wallace, 1965). Those works appeared at the end of what Wherry (1957) referred to as a renaissance period (roughly 1940–1960) "marked by the reapplication of validity principles, earlier applied to testing, to the criteria themselves" (p. 3). Thus, one theme apparent from the present review is that of validation, applied to both sides of a prediction equation, and its increasing sophistication over the history of criterion measurement.

Despite persistent arguments that criteria cannot or should not be validated (e.g., Anastasi, 1950; Astin, 1964; Barrett & Kernan, 1987; Freyd, 1923; Lopez, 1968; Ronan & Prien, 1966), we believe that any measure, whether it occurs on one side of a regression equation or another, is capable of validation. We agree with the logic of Banks and Roberson (1985), who maintained that appraisal instruments should be developed as much as tests are. The crucial concept in validation is understanding the measure and its latent construct, which leads to prediction as a by-product and not as a terminal goal.

Despite at least one and often several exhortations per decade (in chronological order: Bingham, 1926; Viteles, 1932; Toops, 1944; Wherry, 1957; Wallace, 1965; Smith, 1976; Landy & Farr, 1983; Borman, 1991), there has been a chronic lack of attention to the conceptual and psychometric characteristics of criteria. This neglect is manifested in the history of construct validation itself, which was developed for predictors that lacked independent criteria (e.g., clinical diagnostic measures). Even though constructs on both sides of a prediction equation can be related to a psychological domain (Binning & Barrett, 1989), criteria were the orphans of the validation process. Both classes of constructs can and should be accorded equivalent status (Patterson, 1946). As general evidence of favoritism toward predictors, note the temporal lag between the development of construct validity (Cronbach & Meehl, 1955) and its application to criteria (e.g., Guion, 1965; James, 1973). At the risk of sounding rhetorical, what measure is more likely to lack an "independent standard" than a criterion measure?

A more specific sign of relative deprivation is the distributions of reliabilities for predictors and criteria. On the basis of review of validity study characteristics by Schmidt, Hunter, and associates, hypothetical reliability distributions for correcting predictors center on .80, whereas the corresponding mean for criterion reliability is .60. Further specific evidence was developed by Lent et al. (1971) in their summary of the *Validity Information Exchange* (VIE; 1954–1963) from *Personnel Psychology.* They found that the bulk of the reported validation studies were conducted using a single criterion (344 of 406). In addition to attenuating relationships in individual studies, Paese and Switzer (1988) showed through simulation the effects that different criterion reliability distributions can have on error rates for validity generalization.

James's (1973) application of construct validation to criteria was a breakthrough, but a lack of attention to it pervades I-O psychology. James argued that evaluating criterion measures requires multiple measures and models of individual performance to integrate the measures. On the basis of this logic he proposed an early latent variable model (LVM) by merging multiple criteria with J. P. Campbell et al.'s (1970) person-process-product model of managerial effectiveness. J. P. Campbell et al. (1970) proposed that individual factors, task demands, organizational climate, and training experiences accounted for the bulk of the systematic variation in job behavior, performance, and organizational effectiveness. James (1973), in turn, specified a theoretical model that incorporated predictor and criterion constructs defined through multiple indicators, reflecting an early application of structural equation modeling for I-O psychology. Work on developing and testing multiple models of criteria, as cited earlier, continues this trend. On the other hand, sophisticated analyses must be used cautiously because of a potential for misuse (Austin & Wolfle, 1991; Breckler, 1990; Cliff, 1983).

A number of subthemes emerge from implementing equal status for criterion measures through construct validation. One is the distinction first made by Cronbach (1949) between typical and maximum performance on predictor instruments, which can be extended to criteria through the idea of matching (similar to the idea of specificity matching of behavior and attitudes). A consideration of this categorization for criteria led to a comparison by Sackett, Zedeck, and Fogli (1988). Those researchers found a fairly low correlation between typical (point-of-sale terminal used to measure a standard grocery cart) and maximum (work sample format measurement of the same cart). The correlation also differed between groups defined on job tenure (experienced vs. inexperienced). Schmitt (1989) made the point that many predictors used in selection may not correspond to the types of task/job performance measures used in validation. Scott and Hamner (1975) showed how performance profiles could be manipulated through the use of a subordinate-confederate to evaluate the ability of raters to pick out systematic variance from random fluctuations.

A second subtheme is the recent refocusing of validation research away from a concern with reliability toward understanding bias, which by definition is systematic but irrelevant variance in the criterion measures (Messick, 1989). An expansion of Brogden and Taylor's (1950b) classification system would be useful. Wiley (1991) and Ackerman and Humphreys (1990) both noted this shift, which has implications for criterion measurement. One is Humphreys' conception of systematic heterogeneity as a method for constructing measures. Systematic heterogeneity consists of deliberate attempts to measure a construct using a wide variety of perspectives (Hulin & Humphreys, 1980). Instead of focusing on high internal consistency and homogeneity, which may result in high reliabilities but narrow measures, heterogeneity is used to reduce bias and thereby increase validity (assuming that each item contributes a small amount of construct variance). Extension of this technique to criterion measurement is illustrated in a multisample field study by Roznowski and Hanisch (1989). Their results, in a study of affect–withdrawal relations, indicated superior performance for the heterogeneous criterion composite (labeled "adaptation syndrome") over several homogeneous composites in terms of accountable variance. A second implication is renewed attention to subgroup bias in criterion measurement (Oppler, Campbell, Pulakos, & Borman, 1992; Pulakos, White, Oppler, & Borman, 1989). Pulakos et al. (1989) used Project A data, and found small effects. Oppler et al. (1992) discussed three structural equation approaches to this topic—total association, direct effect, and differential construct—and illustrated their application to Project A data sets. An advantage of these recent studies is that they have incorporated gender and added Hispanics to the standard Black-White comparisons.

Values in Criterion Research

A second theme that emerged from the review is a persistent tension between various constituencies affected by criteria of work performance (Austin et al., 1991). We view this as an issue of values. Management is often focused on the administrative and evaluative aspects of criterion measurement, which facilitates their personnel classification and decision-making roles. As we noted earlier, research purists are often more concerned with the understanding provided by integrated construct validation and substantive research (Schwab, 1980). On the other hand, employees tend to focus on the feedback and developmental functions of criterion measures used in performance appraisal. We can only speculate about the results of such differing perspectives, recalling the problems experienced with

acceptance of the forced-choice method. However, even this three-constituency approach to characterizing competing values may be too simplistic. Within each of these interest groups are embedded several others, each of which has different aims and different means of persuasion to press their claims. For example, the employee group could be divided on the basis of tenure into two or more groups (along a continuum); more senior employees might favor criteria that stress seniority and may appeal to their union so that a performance-based measurement system might be shelved. On the other hand, consider a grouping based on performance levels. Poor-performing employees, regardless of seniority, may favor less sensitive measures. This complexity means that research to investigate the costs and benefits of numerous perspectives is needed to advise organizations how to implement a performance evaluation process that optimizes individual and organizational goal attainment.

If we are to accept the idea that the criterion domain is somehow deficient when omitting individual and societal goals (cf. Bass, 1952), then it seems incumbent on researchers to include them in order to be faithful to the scientific enterprise. However, as the net is cast wider to capture the ultimate criterion, another problem arises that must be faced and dealt with directly. That is, if we assume that each of these criteria are in turn multidimensional, how do we deal with the multidimensionality of subcriteria in practice? Should they be represented through a single score, statistically or rationally weighted, or should they each have a unique prediction equation? In either case, tradeoffs are involved that are anathema to scientists and practitioners alike. One provisional recommendation is that preferences regarding system/product/service performance need to be derived from individuals who represent positions that derive value from the organization's outputs (customers or clients; e.g., workers, supervisors, management, stockholders, consumers, and possibly the public-at-large). Those values need to be somehow hierarchically ordered and weighted with a subset used to identify potentially critical aspects of performance (criteria) for use in evaluating individual and organizational effectiveness. Measures of criteria can be combined into a composite representative of the weighted policies or profiles of value elicited from customers. Drawing on the conceptual work of Fiske (1951) and procedures used in program evaluation research, Villanova (1992) describes an approach to deriving criteria using links between customers, job function relevance, value dimension importance, and our more restrictive definition of criteria. A significant idea represented in this work is the need to disaggregate values maintained by different customers and to identify the relevance of performance on specific job functions for satisfying customer requirements. Additional work is needed on how to identify suitable measures of criteria and ways to incorporate the diverse goals of different customers into indices of effectiveness, as well as the incremental utility implications of developmental activities.

Earlier we cited the work of Zammuto (1984) and Keeley (1984) as relevant to the incorporation of values in criterion research. Additional related work on this topic appears in the educational evaluation literature under the title "stakeholder" and "stakeholder based evaluation" (Bryk, 1983; Greene, 1988). Insights from research based on those conceptions may provide guidance for I-O psychologists interested in incorporating constituency values into criterion development and evaluation (cf. Henry, Dickey, & Areson, 1991).

There is another issue, which appears to break down along theoretical versus practical/legal lines, as to whether organizations can permissibly evaluate extra-role behaviors, and thus cast such a wide net over employee behavior. If there is some sort of implicit psychological or explicit legal contract governing the relationship between employee and employer, is it legal to evaluate on extra-role behaviors, though they may be of value to potential customers? Theoretically, the answer would seem to be yes in the service of investigating a construct defined as "value to the organization," but practically and legally, the opposite might hold.

Research–Practice Interface

When contemplating the history of criterion measurement, we noticed a tendency for research and practice to approximate parallel trains. That is, both trains are heading in the same direction and are within sight of each other, but they are not converging despite the illusion caused by the horizon. According to Bretz et al. (1992), there has been uneven integration between research and practice across the topics they reviewed. There is probably no ready explanation for why this has occurred, but rating format and cognitive research provide two examples of how research does not always inform practice. In the former case, quantitative-psychometric standards are the preferred means of evaluating formats, while there has been less use of the classes of standards termed qualitative and utilization (Jacobs et al., 1980). One way to extend research in the format area might be to examine the relationship between the three categories of quantitative, utilization, and qualitative criteria for criteria. Specifically, do different psychometric characteristics influence reactions to appraisals by raters, ratees, or both? Similarly, the cognitive paradigm as applied to performance ratings has concentrated on laboratory methods and student samples that restrict generalization (cf. Ilgen & Favero, 1985). If we adopt the role-context-vehicle model that Landy and Farr presented in 1980, ratings, once generated, should serve as input into a human resource management system, at a minimum including links to feedback, training, goal-setting, reward, and career components. To date, there have been few efforts even to describe such fit, with a few exceptions. Bernardin (1986), in discussing a comprehensive performance appraisal system, warned researchers that many assumptions (he listed nine) of traditional approaches served as obstacles to effectiveness. He then identified 15 relevant parameters and provided a flow-chart to guide decision-making in system development. Finally, Bernardin's (1986) last sentence was prescient. He concluded, "Perhaps this single factor, holding raters accountable for their ratings, just as they are held accountable for the administration of other expensive organizational resources, will do more to improve the effectiveness of a PA system than any other technique or intervention that could be recommended" (p. 302). Subsequent work by Klimoski and Inks (1990) showed that accountability forces could influence rating quality.

Recommendations for Future Research

Several research recommendations become apparent from a historical examination of the criterion problem. It may be profitable to pay attention to related areas of psychology that encounter criterion problems. One such area is developmental psychology. Recall that Smith and Kendall (1963) used the Fels Child/Parent Behavior rating scale format to develop the BARS format. Surber (1984) reviewed the use of rating scales in developmental research, harkening back to the work of Weiss (1933). A promising alliance seems to be development of explanations for dynamic criteria. A second example of an alliance is provided by Jaccard (1979), an attitude researcher, who analyzed and classified behavioral measures within a Fishbein–Ajzen attitudinal framework to produce a "general theory of the single act criterion" (p. 78). Starting from a behavior by occasion matrix, Jaccard defined single/multiple act and single/repeated observation behavioral data within homogeneous or heterogeneous conditions. Jaccard's facets could be combined with Kane's (1986) work on performance distribution assessment, which provides a framework for analyzing performance variability over multiple acts (enabling also the analysis of dynamic criterion hypotheses). The merger would extend Jaccard's single-act model through the use of a mechanism for modeling variability in performance over multiple acts, rather than focusing only on level of behavior or results of behavior. A final example from social psychology is the work of Berman and Kenny (1976), which manipulated halo in a laboratory setting. This research suggested that the problem of halo could be attacked from a different perspective.

A further suggestion is to examine the human engineering domain for insights and translations. Ilgen and Schneider (1991) reviewed performance and its measurement using human factors and production/operations management concepts as well as the traditional human resource management focus on individuals. The focus of their review, across the disciplines, was on five aspects of performance measurement: purpose, source, criteria, standards for performance measures, and methods of measurement. Earlier work by Chiles (1967), Uhlaner (1972; Uhlaner & Drucker, 1964, 1980), and Meister (1985) is often ignored by I-O researchers, but it provides a useful emphasis on person-machine and system concepts that are becoming increasingly useful with work place technology innovations (i.e., Workplace 2000; Johnston & Packer, 1987). Meister (1985) and Uhlaner (1972) presented several implications of a systems focus for criteria. A final suggestion is to investigate performance within educational systems, which permit the study of performance at individual (students as service recipients, teachers as service providers) and higher levels (class or school district). Ryans's (1960) study of teacher characteristics and Biddle and Ellena's (1964) edited collection of work on teacher effectiveness could be reexamined for research questions. Also of interest are multilevel approaches (Bock, 1989; Bryk & Raudenbush, 1987), which complement methods developed within I-O psychology (Dansereau, Alutto, & Yammarino, 1984).

A related point is that work performance, if defined as transactional between people and environments in Gordon Allport's terms (cf. Astin, 1964; Hanges et al., 1990), may be more situationally specific than KSAOs used as predictors. Psychologists have been criticized for a failure to study the situation (cf. Barker, 1968; Moos, 1976; Sells, 1963). If such is the case, searching for general performance constructs may be futile and more refined taxonomies will be required to match criterion measures or composites to specific situations. It may not be possible to match a general ability factor with a corresponding general performance factor.

The problems of criteria have been and remain multiple. Criteria are dynamic, multidimensional, situation-specific, and serve multiple functions. Given this complexity, it would be unreasonable to assume that a hypothetico-deductive model of explanation, which provides premises from which events can be deduced, could offer a sufficient basis for understanding criteria. Similarly, the pattern model described by Kaplan (1964), which explains events on the basis of their relationships to other events in a network, also seems insufficient in this regard. The pattern model requires that the observations be interpretable within and constitutive of a larger system. The wider the system, and the more reproducible it is in different contexts, the greater the subsequent understanding of and ability to explain it. However, the reproducibility of the system depends on the situation, the values used in selecting and measuring criteria, and the dynamic nature of criteria, all of which contribute to criterion dimensionality and system instability. Cattell's (1988b) inductive-hypothetico-deductive (IHD) model, in the form of an iterated spiral, appears appropriate but lacks adherents and principles at this time. For these reasons, the criterion problem is likely to remain a problem for those uncomfortable with probabilistic, partial, and conditional explanations. Those who seek determinate solutions to the criterion problem are cautioned that "finality is the mark of runic explanations, not scientific ones; the road of inquiry is always open, and it reaches beyond the horizon" (Kaplan, 1964, p. 341).

References

Ackerman, P., & Humphreys, L. G. (1990). Individual differences theory in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 223–282). Palo Alto, CA: Consulting Psychologists' Press.

Adkins, D. (1947). *Construction and analysis of achievement tests.* Washington, DC: U.S. Government Printing Office.

Allen, F. L. (1931). *Only yesterday: An informal history of the nineteen-twenties.* New York: Harper Brothers.

Alliger, G. M., & Hosoda, M. (1992, May). *The shape of work: The distribution of measures of work performance, and implications for Total Quality Management.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Montreal, Canada.

Alvares, K. M., & Hulin, C. L. (1972). Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. *Human Factors, 14,* 295–308.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological tests.* Washington, DC: American Psychological Association.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: Author.

American Psychological Association. (1989). In the Supreme Court of the United States: *Clara Watson vs. Fort Worth Bank and Trust. American Psychologist, 43,* 1019–1028.

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement, 10,* 67–78.

Anderson, N. H., & Shanteau, J. (1977). Weak inference with linear models. *Psychological Bulletin, 84,* 1155–1170.

Anderson, R. N. (1929–1930). Measurement of clerical ability: A critical review of proposed tests. *Journal of Personnel Research, 8,* 232–244.

Angell, J. R. (1911). Philosophical and psychological usage of the terms mind, consciousness, and soul. *Psychological Bulletin, 8,* 46–47.

Astin, A. W. (1964). Criterion-centered research. *Educational and Psychological Measurement, 24,* 807–822.

Atkin, R. S., & Conlon, D. J. (1978). Behaviorally anchored rating scales: Some theoretical issues. *Academy of Management Review, 3,* 119–128.

Austin, J. T., Humphreys, L. G., & Hulin, C. L. (1989). Another view of dynamic criteria: A critical reanalysis of Barrett, Caldwell, and Alexander. *Personnel Psychology, 42,* 583–596.

Austin, J. T., Villanova, P., Kane, J. S., & Bernardin, H. J. (1991). Construct validation of performance measures: issues, development, and evaluation of indicators. In G. R. Ferris & K. M. Rowland (Eds.), *Research in personnel and human resources management* (Vol. 9, pp. 159–233). Greenwich, CT: JAI Press.

Austin, J. T., & Wolfle, L. M. (1991). Annotated bibliography of structural equation modelling: Technical work. *British Journal of Mathematical and Statistical Psychology, 44,* 93–152.

Baggaley, A. (1974). A scheme for classifying rating methods. *Personnel Psychology, 27,* 139–144.

Bailey, C. T. (1983). *The measurement of job performance.* Aldertshot, England: Gower Press.

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

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology, 38,* 335–345.

Banks, C. G., & Roberson, L. (1985). Performance appraisers as test developers. *Academy of Management Review, 10,* 128–142.

Banner, D. K., & Cooke, R. A. (1984). Ethical dilemmas in performance appraisal. *Journal of Business Ethics, 3,* 327–333.

Bare, R. H. (1954). *Bias as related to rater contacts.* Unpublished doctoral dissertation, The Ohio State University, Columbus.

Barker, R. G. (1968). *Ecological psychology.* Palo Alto, CA: Stanford University Press.

Barnes, R. M. (1940). *Motion and time study* (2nd ed.). New York: Wiley.

Barrett, G. V. (1972). Symposium: Research models of the future for industrial and organizational psychology. 1. Introduction. *Personnel Psychology, 25,* 1–17.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1985). The concept of dynamic criteria: A critical reanalysis. *Personnel Psychology, 38,* 41–56.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1989). The predictive stability of ability requirements for task performance: A critical reanalysis. *Human Performance, 2,* 167–181.

Barrett, G. V., & Kernan, M. (1987). Performance appraisal and terminations: A review of court decisions since *Brito v. Zia* with implications for personnel practices. *Personnel Psychology, 40,* 489–503.

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

Bass, B. M. (1952). Ultimate criteria of organizational worth. *Personnel Psychology, 5,* 157–174.

Bechtoldt, H. (1947). Problems in establishing criterion measures. In D. B. Stuit (Ed.), *Personnel research and test development in the Bureau of Naval Personnel* (pp. 357–379). Princeton, NJ: Princeton University Press.

Bechtoldt, H. (1951). Selection. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1237–1266). New York: Wiley.

Becker, B. E., & Cardy, R. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. *Journal of Applied Psychology, 71,* 662–671.

Bellows, R. M. (1941). Procedures for evaluating vocational criteria. *Journal of Applied Psychology, 25,* 499–513.

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

Bergen, G. L. (1934–1935). The practical use of tests in appraising occupational fitness. *Personnel Journal, 13,* 73–81.

Berk, R. A. (Ed.). (1986). *Performance assessment: Methods and applications.* Baltimore: Johns Hopkins University Press.

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

Berman, J. S., & Kenny, D. A. (1976). Correlational bias in observational ratings. *Journal of Personality and Social Psychology, 34,* 263–273.

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

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

Bernardin, H. J. (1986). A performance appraisal system. In R. A. Berk (Ed.), *Performance assessment* (pp. 277–303). Baltimore: Johns Hopkins University Press.

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

Bernardin, H. J., & Beatty, R. (1984). *Performance appraisal: Assessing human behavior at work.* Boston: Kent-PWS.

Bernardin, H. J., & Buckley, M. R. (1981). A consideration of strategies in rater training. *Academy of Management Review, 6,* 205–212.

Bernardin, H. J., Cardy, R. L., & Carlyle, J. J. (1982). Cognitive complexity and appraisal effectiveness: Back to the drawing board? *Journal of Applied Psychology, 67,* 151–160.

Bernardin, H. J., & Cascio, W. F. (1988). Performance appraisal and the law. In R. Schuler, S. A. Youngblood, & V. L. Huber (Eds.), *Personnel and human resource management* (pp. 235–247). St. Paul, MN: West.

Bernardin, H. J., & Kane, J. (1980). A closer look at behavioral observation scales. *Personnel Psychology, 33,* 809–814.

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

Bernardin, H. J., & Pence, E. C. (1980). Rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65,* 60–66.

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales. *Journal of Applied Psychology, 66,* 458–463.

Bernardin, H. J., & Villanova, P. (1986). Performance appraisal. In E. A. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 43–62). Lexington, MA: Lexington Press.

Bernardin, H. J., & Walter, C. S. (1977). Effectiveness of rater training

and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology, 62,* 64–69.

Bezanson, A., Willits, J. H., Chalufour, F., & White, L. (1922). A study in labor mobility. *Annals of the American Academy of Political and Social Science, 103* (Whole No. 192).

Biddle, B. J., & Ellena, W. J. (Eds.). (1964). *Contemporary research on teacher effectiveness.* New York: Holt, Rinehart, & Winston.

Bills, M. A. (1938–1939). Present trends in selection for employment. *Personnel, 15,* 184–217.

Binet, A., & Simon, T. (1908). Le development de intelligence chez les enfants [The development of intelligence among infants]. *L'Annee Psychologique, 14,* 1–94.

Bingham, W. V. (1923). On the possibility of an applied psychology. *Psychological Review, 30,* 289–305.

Bingham, W. V. (1926). Measures of occupational success. *Harvard Business Review, 5,* 1–10.

Bingham, W. V. (1928). Personality and public accidents: The study of accident-prone drivers. *Transactions of the Seventeenth Annual Safety Congress,* 5–13.

Bingham, W. V. (1931). The prone-to-accident driver. *Proceedings of the 17th Annual Conference on Highway Engineering,* 3–12.

Bingham, W. V. (1939). Halo: Valid and invalid. *Journal of Applied Psychology, 23,* 221–228.

Bingham, W. V., & Davis, W. T. (1924). Intelligence tests and business success. *Journal of Applied Psychology, 8,* 1–22.

Bingham, W. V., & Freyd, M. (1926). *Procedures in employment psychology.* Washington, DC: Shaw.

Binning, J., & Barrett, G. V. (1989). Validity of personnel decisions: A review of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478–494.

Bird, N. (1931). Relationships between experience factors, test scores, and efficiency. *Archives of Psychology, 126*(X).

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

Bloom, B. S. (1964). *Stability and change in human characteristics.* New York: Wiley.

Bluedorn, A. C. (1982). A unified model of turnover from organizations. *Human Relations, 35,* 135–153.

Blumberg, M., & Pringle, C. D. (1982). The missing opportunity in organizational research: Some implications for a theory of work performance. *Academy of Management Review, 7,* 560–569.

Bobko, P. (1990). Multivariate correlational analysis. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 637–686). Palo Alto, CA: Consulting Psychologists' Press.

Bock, R. D. (Ed.). (1989). *Multilevel analysis of educational data.* New York: Academic Press.

Boice, R. (1983). Observational skills. *Psychological Bulletin, 93,* 3–29.

Bolanovich, D. J. (1946). Statistical analysis of an industrial rating chart. *Journal of Applied Psychology, 30,* 23–31.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

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

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

Borman, W. C. (1977). Consistency of rating accuracy and rating error in the judgment of human performance. *Organizational Behavior and Human Performance, 20,* 238–252.

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

Borman, W. C. (1979a). Format and training effects on rating accuracy using behavior scales. *Journal of Applied Psychology, 64,* 410–421.

Borman, W. C. (1979b). Individual difference correlates of rating accuracy using behavior scales. *Applied Psychological Measurement, 3,* 103–115.

Borman, W. C. (1986). Behavior-based rating scales. In R. A. Berk (Ed.), *Performance assessment: Methods & applications* (pp. 110–120). Baltimore: Johns Hopkins University Press.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 271–326). Palo Alto: Consulting Psychologists' Press.

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

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology, 76,* 863–872.

Bradshaw, F. F. (1930). The American Council on Education rating scale: Its reliability, validity, and use. *Archives of Psychology, 119.*

Braeman, J., Bremner, R. H., & Brody, D. (Eds.). (1968). *Change and continuity in twentieth-century America: The 1920s.* Columbus: The Ohio State University Press.

Bray, C. W. (1948). *Psychology and military efficiency.* Princeton, NJ: Princeton University Press.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107,* 260–273.

Bretz, R. D., Jr., Milkovich, G. T., & Read, W. (1990). *Comparing the performance appraisal practices in large firms with the directions in research literature: Learning more and more about less and less.* Ithaca, NY: New York State School of Industrial and Labor Relations.

Bretz, R. D., Jr., Milkovich, G. T., & Read, W. (1992). The currect state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management, 18,* 321–352.

Brito v. Zia Co., 478 F.2d 1200 (10th Cir. 1973).

Brogden, H., & Taylor, E. K. (1950a). The dollar criterion: Applying the cost accounting concept to criterion construction. *Personnel Psychology, 3,* 133–154.

Brogden, H., & Taylor, E. K. (1950b). The theory and classification of criterion bias. *Educational and Psychological Measurement, 10,* 159–186.

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

Browne, M. W., & DuToit, S. H. C. (1991). Models for learning data. In L. Collins & J. Horn (Eds.), *Best methods for the analysis of change* (pp. 47–68). Washington, DC: American Psychological Association.

Bryk, A. S. (Ed.). (1983). *Stakeholder-based evaluation* (New Directions in Program Evaluation, No. 17). San Francisco: Jossey-Bass.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–158.

Buckley, M. R., Villanova, P., & Benson, P. G. (1989). Contrast effects in performance ratings: Another look across time. *Applied Psychology: An International Review, 38,* 131–143.

Burt, C., Gaw, F., Ramsey, L., Smith, M., & Spielman, W. (1926). *A study in vocational guidance* (Industrial Fatigue Research Board Report No. 33). London: Her Majesty's Stationery Office.

Burtt, H. E. (1926). *Principles of employment psychology.* Boston: Houghton-Mifflin.

Buxton, C. E. (1985). American functionalism. In C. E. Buxton (Ed.), *Points of view in the history of modern psychology* (pp. 113–140). New York: Academic Press.

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

Campbell, J. P. (1990a). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 687–732). Palo Alto, CA: Consulting Psychologists' Press.

Campbell, J. P. (1990b). An overview of the Army selection and classification project. *Personnel Psychology, 43,* 231–239.

Campbell, J. P., & Campbell, R. J. (Eds.). (1988). *Productivity in organizations.* San Francisco: Jossey-Bass.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology, 57,* 15–22.

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

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology, 43,* 313–333.

Campbell, J. P., & Pritchard, R. D. (1976). Motivation theory in industrial and organizational psychology. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 63–130). Chicago: Rand McNally.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). *An investigation of sources of bias in the prediction of job performance: A six-year study* (PR-73-37). Princeton, NJ: Educational Testing Service.

Campion, M. A. (1991). Meaning and measurement of turnover: Comparison of alternative measures and recommendations for research. *Journal of Applied Psychology, 76,* 199–212.

Carlson, R. E. (1967). Selection interview decisions: The effect of interviewer experience, relative quota situation, and applicant sample on interviewer decisions. *Personnel Psychology, 20,* 259–280.

Carson, K. P., Cardy, R. L., Dobbins, G. H., & Stewart, G. L. (1992, May). *Determinants and domains of performance: Implications for the evaluation of employee performance.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Montreal, Canada.

Cascio, W. F., & Bernardin, H. J. (1981). Implications of performance appraisal litigation for personnel decisions. *Personnel Psychology, 34,* 211–226.

Cattell, R. B. (1964). Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology, 55,* 1–22.

Cattell, R. B. (1988a). The data box: Its ordering of total resources in terms of possible relational systems. In J. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 69–130). New York: Plenum.

Cattell, R. B. (1988b). Psychological theory and scientific method. In J. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 3–20). New York: Plenum.

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology, 23,* 323–331.

Chapman, J. C. (1921). *Army trade tests.* New York: Holt.

Chapman, J. C., & Toops, H. A. (1919). A written trade test: Multiple choice method. *Journal of Applied Psychology, 3,* 358–365.

Chiles, W. D. (1967). Methodology in the assessment of complex performance: Discussion and conclusions. *Human Factors, 9,* 385–392.

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement, 9,* 31–36.

Civil Rights Act of 1964, 42 U.S.C. § 2000e et seq. (1964).

Civil Rights Act of 1991, P. L. 102–166, 105 Stat. 1071 (1991).

Clarke, H. W. (1956). *An experimental investigation of theorems relating to the structure and content of rating instruments.* Unpublished doctoral dissertation, The Ohio State University, Columbus.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74,* 130–135.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research, 18,* 115–126.

Cobb, P. W. (1940). The limit of usefulness of accident rate as a measure of accident proneness. *Journal of Applied Psychology, 24,* 154–159.

Conrad, H. S. (1933). The personal equation in ratings: A systematic evaluation. *Journal of Educational Psychology, 24,* 39–46.

Cook, T. D., & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 223–326). Chicago: Rand McNally.

Cooper, J. H. (1940). Rating forms. In W. H. Stead, C. L. Shartle, & Associates (Eds.), *Occupational counseling techniques* (pp. 49–72). New York: American Book.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90,* 218–244.

Cote, J. A., & Buckley, M. R. (1987). Estimating trait, method, and error variance: Generalizing across seventy construct validation studies. *Journal of Marketing Research, 24,* 315–318.

Creager, J. A., & Harding, F. D., Jr. (1958). A hierarchical factor analysis of foreman behavior. *Journal of Applied Psychology, 42,* 197–203.

Crennan, C. H., & Kingsbury, F. W. (Eds.). (1923). Psychology and industry. *Annals of the American Academy of Political and Social Science, 110* (Whole No. 199).

Cronbach, L. J. (1949). *Essentials of psychological testing.* New York: Harper.

Cronbach, L. J. (1971). Validity. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). New York: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Gleser, G. (1965). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Cronshaw, S. (1989, April). *A philosophical analysis of job performance criteria.* Paper presented at the Fourth Annual Conference of the Society for Industrial and Organizational Psychology, Boston.

Cureton, E. E., Jr. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Dalton, D., & Todor, W. (1979). Turnover turned over: An expanded and positive perspective. *Academy of Management Review, 4,* 225–235.

Dansereau, F., Alutto, J., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The varient approach.* Englewood Cliffs, NJ: Prentice-Hall.

Dansereau, F., & Markham, S. E. (1987). Levels of analysis in personnel and human resources management. In K. Rowland & G. R. Gerris (Eds.), *Research in personnel and human resources management* (Vol. 5, pp. 1–50). Greenwich, CT: JAI Press.

Darley, J. G. (1968). 1917: A journal is born. *Journal of Applied Psychology, 52,* 1–9.

Dawis, R., & Lofquist, L. (1984). *A psychological theory of work adjustment* (2nd ed.). Minneapolis: University of Minnesota Press.

Deadrick, D. L., & Madigan, R. J. (1990). Dynamic criteria revisited: A longitudinal study of performance stability. *Personnel Psychology, 44,* 717–744.

DeCotiis, T. A., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. *Academy of Management Review, 3,* 635–645.

Deming, W. E. (1986). *Out of the crisis.* Cambridge, MA: MIT Institute for Advanced Engineering Study.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33,* 360–396.

DeNisi, A., & Williams, K. J. (1988). Cognitive approaches to performance appraisal. In G. R. Ferris & K. M. Rowland (Eds.), *Research in personnel and human resources management* (Vol. 6, pp. 109–155). Greenwich, CT: JAI Press.

Dennis, W. E. (1954). Predicting scientific productivity in later maturity from records of earlier decades. *Journal of Gerontology, 9,* 465–467.

DeSantis, V. P. (1989). *The shaping of modern America: 1988–1920* (2nd ed.). Arlington Heights, IL: Forum Press.

Dickinson, E. Z. (1937). Validity and independent criteria in tests and ratings. *Journal of Applied Psychology, 21,* 522–527.

Dickinson, T. I., & Tice, T. E. (1973). A multitrait-multimethod analysis of scales developed by translation. *Organizational Behavior and Human Performance, 9,* 421–439.

Dipboye, R. (1985). Some neglected variables in research on discrimination in appraisals. *Academy of Management Review, 10,* 116–127.

Dobbins, G. H., Cardy, R. L., & Carson, K. P. (1991). Examining fundamental assumptions: A contrast of person and system approaches to human resource management. In G. R. Ferris & K. M. Rowland (Eds.), *Research in personnel and human resource management* (Vol. 9, pp. 1–38). Greenwich, CT: JAI Press.

Dobbins, G. H., Cardy, R. L., & Platz-Vieno, S. J. (1990). A contingency approach to appraisal satisfaction: An initial investigation of the joint effects of organizational variables and appraisal characteristics. *Journal of Management, 16,* 619–632.

Dooher, M. J., & Marquis, V. (Eds.). (1950). *Rating employee and supervisory performance.* New York: American Management Association.

Dorcus, R. M., & Jones, M. H. (1950). *Handbook of employee selection.* New York: McGraw-Hill.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 577–636). Palo Alto, CA: Consulting Psychologists' Press.

Dreher, G. F., & Sackett, P. R. (Eds.). (1983). *Perspectives on employee staffing and selection.* Homewood, IL: Irwin.

Dubin, R. (1976). Theory building in applied areas. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 17–39). Chicago: Rand McNally.

Duncan, A. J. (1974). *Quality control and industrial statistics.* Homewood, IL: Irwin.

Dunham, R. B. (1974). Ability-skill relationships: An empirical explanation of change over time. *Organizational Behavior and Human Performance, 12,* 373–382.

Dunnette, M. D. (1963a). A modified model for test validation and research. *Journal of Applied Psychology, 47,* 317–323.

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

Dunnette, M. D. (Ed.). (1976). *Handbook of industrial and organizational psychology.* Chicago: Rand McNally.

English, H. B., & English, A. C. (1958). *A comprehensive dictionary of psychological and psychoanalytic terms.* New York: Longmans, Green & Company.

Ewart, E., Seashore, S. E., & Tiffin, J. (1941). A factor analysis of an industrial merit rating scale. *Journal of Applied Psychology, 25,* 481–486.

Fandt, P. M., & Ferris, G. R. (1990). The management of information and impressions: When employees behave opportunistically. *Organizational Behavior and Human Decision Processes, 45,* 140–158.

Farmer, E. (1921). *Time and motion study* (Industrial Fatigue Research Board Report No. 14). London: His Majesty's Stationery Office.

Farmer, E. (1933). The reliability of the criteria used for assessing the value of vocational tests. *British Journal of Psychology, 24,* 109–119.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66,* 127–148.

Feldman, J. M. (1986). Instrumentation and training for performance appraisal: A perceptual cognitive view. In K. M. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 4, pp. 45–99). Greenwich, CT: JAI Press.

Ferguson, L. W. (1950). The L.O.M.A. merit rating scales. *Personnel Psychology, 3,* 193–216.

Ferguson, L. W. (1962–1965). *The heritage of industrial psychology.* Hartford, CT: Finlay Press.

Fichman, M. (1984). A theoretical approach toward understanding employee absence. In P. S. Goodman & R. Atkin (Eds.), *Absenteeism: New approaches to understanding, measuring, and managing employee absence* (pp. 1–46). San Francisco: Jossey-Bass.

Fichman, M. (1988). Motivational consequences of absence and attendance: Proportional hazard estimation of a dynamic motivation model. *Journal of Applied Psychology, 73,* 119–134.

Fichman, M. (1989). Attendance makes the heart grow fonder: A hazard rate approach to modeling attendance. *Journal of Applied Psychology, 74,* 325–335.

Fisher, C. D. (1989). Current and recurrent challenges in HRM. *Journal of Management, 15,* 157–180.

Fiske, D. W. (1951). Values, theory, and the criterion problem. *Personnel Psychology, 4,* 93–98.

Flanagan, J. C. (Ed.). (1947). *The aviation psychology program in the Army Air Forces* (19 vols.). Washington, DC: U.S. Government Printing Office.

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

Flanagan, J. C. (1956). The evaluation of methods in applied psychology and the problem of criteria. *Occupational Psychology, 30,* 1–9.

Flanagan, J. C., Fiske, D. W., Bass, B. M., Carter, L. F., & Kelly, E. L. (1954). Situational performance tests (A symposium). *Personnel Psychology, 7,* 461–497.

Fleishman, E. A. (1975). Toward a taxonomy of human performance. *American Psychologist, 30,* 1127–1149.

Fleishman, E. A., & Mumford, M. (1989). Abilities as causes of individual differences in skill acquisition. *Human Performance, 2,* 201–223.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance.* New York: Academic Press.

Fogli, L., Hulin, C. L., & Blood, M. R. (1971). Development of first-level behavioral job criteria. *Journal of Applied Psychology, 55,* 3–8.

Ford, J. K., Kraiger, K., & Schechtman, S. L. (1986). Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. *Psychological Bulletin, 99,* 330–337.

Freeberg, N. E. (1955). *The effect of relevant contact upon the validity of ratings.* Unpublished doctoral dissertation, The Ohio State University, Columbus.

Freeberg, N. E. (1976). Criterion measures for youth-work training programs: The development of relevant performance dimensions. *Journal of Applied Psychology, 61,* 537–545.

Freyd, M. (1923). The graphic rating scale. *Journal of Educational Psychology, 14,* 83–102.

Freyd, M. (1923–1924). Measurement in vocational selection: An outline of research procedure. *Journal of Personnel Research, 2,* 215–249, 268–284, 377–385.

Freyd, M. (1926). What is applied psychology? *Psychological Review, 33,* 308–314.

Fromkin, H. L., & Streufert, S. (1976). Laboratory experimentation. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 415–465). Chicago: Rand McNally.

Fryer, D. H. (1922). Occupational intelligence standards. *School and Society, 16,* 273–277.

Fryer, D. H. (1934–1935). Intelligence tests in industry. *Personnel Journal, 13,* 321–323.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75–90.

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. *Psychological Review, 67,* 347–356.

Geissler, L. R. (1917). What is applied psychology? *Journal of Applied Psychology, 1,* 46–60.

Ghiselli, E. E. (1956). Dimensional problems of criteria. *Journal of Applied Psychology, 40,* 1–4.

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

Gilbreth, F. B. (1909). *Bricklaying system.* Chicago: Myron C. Clark.

Gilbreth, F. B. (1911). *Motion study.* New York: Van Nostrand.

Gilbreth, F. B., & Gilbreth, L. M. (1916). *Fatigue study.* New York: Sturgis & Walton.

Gilbreth, F. B., & Gilbreth, L. M. (1924). The efficiency engineer and the industrial psychologist. *Journal of the National Institute of Industrial Psychology, 2,* 40–45.

Giles, W. F., & Mossholder, K. W. (1990). Employee reactions to contextual and session components of performance appraisal. *Journal of Applied Psychology, 75,* 371–377.

Glickman, A. S. (1955). Effects of negatively skewed ratings on motivations of the rated. *Personnel Psychology, 8,* 39–47.

Goodenough, F. L. (1949). *Mental testing.* New York: Rinehart.

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

Green, B. F., Jr., & Hall, J. F. (1984). Quantitative methods for literature reviews. *Annual Review of Psychology, 35,* 37–53.

Greenberg, J. (1986). Determinants of perceived fairness of performance evaluations. *Journal of Applied Psychology, 71,* 340–342.

Greene, J. C. (1988). Stakeholder participation and utilization in program evaluation. *Evaluation Review, 12,* 91–116.

Greene, L., Bernardin, H. J., & Abbott, J. (1985). A comparison of rating formats after correction for attenuation. *Educational and Psychological Measurement, 45,* 503–509.

Greenwood, M., & Woods, H. M. (1919). *The incidence of industrial accidents, with special reference to multiple accidents* (Industrial Fatigue Research Board Report No. 4). London: His Majesty's Stationery Office.

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

Grings, W. W. (1952). The evaluation of experimentally controlled criteria. *Psychological Bulletin, 49,* 333–338.

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

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

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

Guion, R. M. (1976). Recruiting, selection, and job placement. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 777–828). Chicago: Rand McNally.

Guion, R. M. (1979). *Principles of work sample testing: III. Construction and evaluation of work sample tests* (ARI-TR-79-A10). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11,* 385–398.

Guion, R. M. (1983). Comments on Hunter. In F. J. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 267–275). Hillsdale, NJ: Erlbaum.

Guion, R. M. (1991). Personnel assessment, selection, and placement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 328–397). Palo Alto, CA: Consulting Psychologists' Press.

Hackett, J. D. (1928–1929). Rating legislators. *Personnel Journal, 7,* 130–131.

Hackett, R. D., Bycio, P., & Guion, R. M. (1989). Absenteeism among hospital nurses: An idiographic longitudinal analysis. *Academy of Management Journal, 32,* 424–453.

Hale, M., Jr. (1982). History of employment testing. In A. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (pp. 3–38). Washington, DC: National Academy Press.

Hall, G. S., Baird, J., & Geissler, L. R. (1917). Foreward. *Journal of Applied Psychology, 1,* 5–7.

Hammer, T. H., & Landau, J. C. (1981). Methodological issues in the use of absence data. *Journal of Applied Psychology, 66,* 574–581.

Haney, W. (1981). Validity, vaudeville, and values: A short history of social concerns over standardized testing. *American Psychologist, 36,* 1021–1034.

Hanges, P. M., Schneider, B. J., & Niles, K. (1990). Stability of performance: An interactionist perspective. *Journal of Applied Psychology, 75,* 658–667.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-superior, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41,* 43–62.

Harrison, D. A., & Hulin, C. L. (1989). Investigations of absenteeism: Using event history models to study the absence-taking process. *Journal of Applied Psychology, 74,* 300–316.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology, 73,* 68–73.

Heidbreder, E. (1933). *Seven psychologies.* New York: Century.

Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology, 39,* 811–826.

Heneman, R. L., Wexley, K. N., & Moore, M. L. (1987). Performance rating accuracy: A critical review. *Journal of Business Research, 15,* 431–448.

Henry, G. T., Dickey, K. C., & Areson, J. C. (1991). Stakeholder participation in educational performance evaluation monitoring systems. *Educational Evaluation and Policy Analysis, 13,* 177–188.

Henry, R. A., & Hulin, C. L. (1987). Stability of skilled performance across time: Some generalizations and limitations on utilities. *Journal of Applied Psychology, 72,* 457–462.

Hobson, C. J., & Gibson, F. W. (1983). Policy capturing as an approach to understanding and improving performance appraisal: A review of the literature. *Academy of Management Review, 8,* 640–649.

Hofmann, D. A., Jacobs, R., & Baratta, J. (in press). Dynamic criteria and the measurement of change. *Journal of Applied Psychology.*

Hofmann, D. A., Jacobs, R., & Gerras, S. J. (1992). Mapping individual performance over time. *Journal of Applied Psychology, 77,* 185–195.

Hollander, E. (1954). Buddy ratings: Military research and industrial implications. *Personnel Psychology, 7,* 385–393.

Hollingworth, H. L. (1920). *Vocational psychology.* New York: Appleton.

Hollingworth, H. L., & Poffenberger, A. T. (1924). *Applied psychology* (2nd ed.). New York: Appleton.

Holt, R. R., & Luborsky, L. (1958). *Personality patterns of psychiatrists* (Vol. 1). New York: Basic Books.

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology, 63,* 579–588.

Hoppock, R. (1936). Ambitious America. *Occupations, 14,* 917.

Horst, P. (Ed.). (1941). *The prediction of personal adjustment* (SSRC Bulletin No. 48). New York: Social Science Research Council.

Hothersall, D. (1990). *History of psychology* (2nd ed.). New York: McGraw-Hill.

Hulin, C. L. (1963). Relevance and equivalence in criterion measures of executive success. *Journal of Industrial Psychology, 1,* 67–78.

Hulin, C. L. (1991). Adaptation, persistence, and commitment in organizations. In M. Dunnette & L. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 445–505). Palo Alto, CA: Consulting Psychologists' Press.

Hulin, C. L., Henry, R., & Noon, S. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relations. *Psychological Bulletin, 107,* 328–340.

Hulin, C. L., & Humphreys, L. G. (1980). Foundations of test theory. In *Construct validity in psychological measurement: Proceedings of a colloquium on theory and application* (pp. 5–12). Princeton, NJ: Educational Testing Service.

Hulin, C. L., & Rousseau, D. M. (1980). Analyzing infrequent events: Once you find them, your troubles begin. *New Directions for Methodology of Social and Behavioral Sciences, 6,* 65–75.

Hull, C. L. (1928). *Aptitude testing.* Yonkers, NY: World Book.

Hull, C. L. (1952). *A behavior system.* New Haven, CT: Yale University Press.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisory ratings. In F. J. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 257–266). Hillsdale, NJ: Erlbaum.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340–362.

Husband, R. W. (1936). In defense of scientific vocational guidance. *Journal of Applied Psychology, 20,* 586–590.

Ilgen, D. R., & Favero, J. (1985). Limits in generalization from psychological research to performance appraisal process. *Academy of Management Review, 10,* 311–321.

Ilgen, D. R., & Feldman, J. (1983). Performance appraisal: A process focus. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 5, pp. 141–196). Greenwich, CT: JAI Press.

Ilgen, D. R., & Hollenback, J. H. (1977). The role of job satisfaction in absence behavior. *Organizational Behavior and Human Performance, 19,* 148–161.

Ilgen, D. R., & Schneider, J. (1991). Performance measurement: A multi-discipline view. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology 1991* (Vol. 6, pp. 71–108). Chichester, England: Wiley.

Inn, A., Hulin, C. L., & Tucker, L. R. (1972). Three sources of criterion variance: Static dimensionality, dynamic dimensionality, and individual dimensionality. *Organizational Behavior and Human Performance, 8,* 58–83.

International Labour Office. (1960). Current trends in industrial psychology. *International Labour Review, 82,* 572–595.

Ivancevich, J. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology, 64,* 502–508.

Jaccard, J. (1979). Personality and behavioral prediction: An analysis of behavioral criterion measures. *New Directions for Methodology of Social and Behavioral Science, 2,* 73–91.

Jacobs, R. R. (1986). Numerical rating scales. In R. A. Berk (Ed.), *Performance assessment: Methods & applications* (pp. 82–99). Baltimore: Johns Hopkins University Press.

Jacobs, R. R., Khafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psychology, 33,* 595–640.

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

Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology, 10,* 93–98.

Johns, G. (1991). Substantive and methodological constraints on behavior and attitudes in organizational research. *Organizational Behavior and Human Decision Processes, 49,* 80–104.

Johnston, W. B., & Packer, A. E. (1987). *Workforce 2000.* Indianapolis, IN: Hudson Institute.

Jones, M. H. (1950). The adequacy of employee selection reports. *Journal of Applied Psychology, 34,* 219–224.

Jordan, A. M. (1923). The validation of intelligence tests. *Journal of Educational Psychology, 14,* 348–366, 414–428.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202.

Kalleberg, A. L., & Kluegel, J. R. (1975). Analysis of the multitrait–multimethod matrix: Some limitations and an alternative. *Journal of Applied Psychology, 60,* 1–9.

Kane, J. S. (1986). Performance distribution assessment. In R. A. Berk (Ed.), *Performance assessment* (pp. 237–273). Baltimore: Johns Hopkins University Press.

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

Kane, J. S., & Lawler, E. E., Jr. (1979). Performance appraisal effectiveness: Its assessment and determinants. In B. M. Staw (Ed.), *Research in organizational behavior* (Vol. 1, pp. 425–478). Greenwich, CT: JAI Press.

Kaplan, A. (1964). *The conduct of inquiry.* San Francisco: Chandler.

Karger, D. W., & Bayha, F. H. (1987). *Engineered work measurement* (4th ed.). New York: Industrial Press.

Katz, R. (1980). Time and work: Toward an integrated perspective. In B. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 2, pp. 81–127). Greenwich, CT: JAI Press.

Katzell, R. A., & Austin, J. T. (in press). From then to now: The development of industrial-organizational psychology in the U.S.A. *Journal of Applied Psychology.*

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

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin, 75,* 34–39.

Keeley, M. (1984). Impartiality and participant-interest theories of organizational effectiveness. *Administrative Science Quarterly, 29,* 1–25.

Kelley, T. L. (1919). Principles underlying the classification of men. *Journal of Applied Psychology, 3,* 50–67.

Kelly, E. L. (1957). Multiple criteria of medical education and their implications for selection. *Journal of Medical Education, 32,* 185–198.

Kelly, J. R., & McGrath, J. E. (1988). *On time and method.* Newbury Park, CA: Sage.

Kelly, P. R. (1958). Reappraisal of appraisals. *Harvard Business Review, 36*(3), 59–68.

Kemery, E., Dunlap, J., & Bedeian, A. G. (1989). The employee separation process: Criterion-related issues associated with tenure and turnover. *Journal of Management, 15,* 417–424.

Kingsbury, F. A. (1922). Analyzing ratings and training raters. *Journal of Personnel Research, 1,* 377–382.

Kingsbury, F. A. (1925–1926). Making rating scales work. *Journal of Personnel Research, 4,* 1–6.

Kingstrom, P., & Bass, B. M. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. *Personnel Psychology, 34,* 263–289.

Kitson, H. D. (1925). *The psychology of vocational adjustment.* Philadelphia: Lippincott.

Klimoski, R., & Inks, L. (1990). Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes, 45,* 194–208.

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

Knauft, E. B. (1947). A classification and evaluation of personnel rating methods. *Journal of Applied Psychology, 31,* 617–625.

Knight, F. B., & Franzen, R. H. (1922). Pitfalls in rating schemes. *Journal of Educational Psychology, 13,* 204–213.

Komaki, J., Collins, R. L., & Temlock, S. (1987). An alternative performance measurement approach: Applied operant measurement in the service sector. *Applied Psychology: An International Review, 36,* 71–89.

Komorita, S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement, 24,* 987–995.

Kornhauser, A. W. (1922). The psychology of vocational selection. *Psychological Bulletin, 19,* 192–229.

Kornhauser, A. W. (1923–1924). A statistical study of a group of specialized office workers. *Journal of Personnel Research, 2,* 103–123.

Kornhauser, A. W. (1926–1927). What are rating scales good for? *Journal of Personnel Research, 5,* 189–193, 309–317, 338–344, 440–446.

Kornhauser, A. W., & Kingsbury, F. W. (1923). *Psychological tests in business.* Chicago: University of Chicago Press.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology, 70,* 56–65.

Krathwohl, D. R. (1985). *Social and behavioral science research.* San Francisco: Jossey-Bass.

Laabs, G. J., & Baker, H. G. (1989). Selection of critical tasks for Navy job performance measures. *Journal of Military Psychology, 1,* 3–16.

Lahey, M. A., & Saal, F. E. (1981). Evidence incompatible with a cognitive compatibility theory of rating behavior. *Journal of Applied Psychology, 66,* 706–715.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41,* 1183–1192.

Landy, F. J. (1992, April). *Waltzing with giants.* Master tutorial presented at the Seventh Annual Conference of the Society for Industrial and Organizational Psychology, Montreal, Canada.

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

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance.* New York: Academic Press.

Landy, F. J., & Rastegary, H. (1989). Criteria for selection. In M. Smith & I. Robertson (Eds.), *Advances in selection and assessment* (pp. 47–65). Chichester, England: Wiley.

Landy, F. J., Vance, R. J., Barnes-Farrell, J. L., & Steele, J. W. (1980). Statistical control of halo error in performance ratings. *Journal of Applied Psychology, 65,* 177–180.

Landy, F. J., & Vasey, J. J. (1984). Theory and logic in human resource management. In K. M. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 2, pp. 1–34). Greenwich, CT: JAI Press.

Landy, F. J., Vasey, J. J., & Smith, F. D. (1984). Methodological problems and strategies in predicting absence. In P. Goodman & R. S. Atkin (Eds.), *Absenteeism: New approaches to understanding, measuring, and managing employee absence* (pp. 110–157). San Francisco: Jossey-Bass.

Landy, F. J., & Zedeck, S. (1983). Introduction. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 1–7). Hillsdale, NJ: Erlbaum.

Landy, F. J., Zedeck, S., & Cleveland, J. (Eds.). (1983). *Performance measurement and theory.* Hillsdale, NJ: Erlbaum.

Latham, G. P., Erez, M., & Locke, E. A. (1988). Resolving scientific disputes by joint design of crucial experiments by the antagonists. *Journal of Applied Psychology, 73,* 753–772.

Latham, G. P., Fay, C., & Saari, L. M. (1980). BOS, BES, and baloney: Raising Kane with Bernardin. *Personnel Psychology, 33,* 815–821.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30,* 255–269.

Latham, G. P., Wexley, K. N., & Pursell, E. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60,* 550–555.

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

Lawshe, C. H., Jr. (1948). *Principles of personnel testing.* New York: McGraw-Hill.

Lee, B. A. (1990). Subjective employment practices and disparate impact: Unresolved issues. *Employee Relations Law Journal, 15,* 403–417.

Lent, R. H., Aurbach, H. A., & Levin, L. S. (1971). Research design and validity assessment. *Personnel Psychology, 24,* 247–274.

Lewin, K. (1931). The conflict between Aristotelian and Galilean modes of thought in contemporary psychology. *Journal of General Psychology, 5,* 141–177.

Lewin, K. (1936). Psychology of success and failure. *Occupations, 13,* 926–930.

Link, H. C. (1918). An experiment in employment psychology. *Psychological Review, 25,* 116–127.

Link, H. C. (1919). *Employment psychology.* New York: Macmillan.

Link, H. C. (1920). The applications of psychology to industry. *Psychological Bulletin, 17,* 335–346.

Lissitz, R., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60,* 1–10.

Locke, E. A. (1982). The ideas of Frederick W. Taylor: An evaluation. *Academy of Management Review, 7,* 14–24.

Locke, E. A. (Ed.). (1986). *Generalizing from laboratory to field settings.* Lexington, MA: Lexington Press.

Longenecker, C. O., Sims, H. P., & Gioia, D. (1987). Behind the mask: The politics of employee appraisal. *Academy of Management Executive, 1,* 183–191.

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

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70,* 66–71.

Lorge, I. (1933). The prediction of vocational success. *Personnel Journal, 12,* 189–197.

Lorge, I. (1936). Criteria for guidance. *Occupations, 14,* 958–962.

Mabe, P. A., III, & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67,* 280–296.

Mace, C. A. (1935). *Incentives: Some experimental studies* (Report 72). London: Industrial Health Research Board.

Macrae, A. (1931). A follow-up of vocationally advised cases. *Journal of the National Institute of Industrial Psychology, 5,* 242–247.

Macrae, A. (1934). Professor Thorndike on vocational guidance. *The Human Factor, 8,* 205–219.

Mahler, W. R. (1947). *Twenty years of merit rating: 1926–1946.* New York: Psychological Corporation.

Mahoney, T. A. (1988). Productivity defined: The relativity of efficiency, effectiveness, and change. In J. P. Campbell & R. J. Campbell (Eds.), *Productivity in organizations* (pp. 13–39). San Francisco: Jossey-Bass.

Maier, N. R. F. (1955). *Psychology in industry* (2nd ed.). Boston: Houghton-Mifflin.

Marble, S. D. (1942). A performance basis for employee evaluation. *Personnel, 18,* 217–226.

March, J. G., & Simon, H. (1958). *Organizations.* New York: Wiley.

Marquis, D. (1944). The mobilization of psychologists for war service. *Psychological Bulletin, 41,* 469–475.

Marsh, S., & Perrin, F. A. C. (1924–1925). An experimental study of the rating scale technique. *Journal of Abnormal and Social Psychology, 19,* 383–399.

Marx, M. H. (Ed.). (1963). *Theories in contemporary psychology.* New York: Macmillan.

McGregor, D. (1957). An uneasy look at performance appraisal. *Harvard Business Review, 35*(5), 89–94.

Meier, N. C. (1943). *Military psychology.* New York: Harper.

Meister, D. (1985). *Behavioral analysis and measurement methods.* New York: Wiley.

Meredith, G. P. (1953). Theory of the 'therblig.' *Occupational Psychology, 27,* 128–136.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55,* 107–122.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Miceli, M. P., Jung, I., Near, J. P., & Greenberger, D. B. (1991). Predictors and outcomes of reactions to pay-for-performance. *Journal of Applied Psychology, 76,* 508–521.

Miner, J. B. (1917). Evaluation of a method for finely graduated estimates of abilities. *Journal of Applied Psychology, 1,* 123–133.

Miner, J. B. (1988). Development and application of the rated ranking technique in performance appraisal. *Journal of Occupational Psychology, 61,* 291–305.

Mitchell, T. R. (1983). The effects of social, task, and situational factors on motivation, performance, and appraisal. In F. J. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 29–59). Hillsdale, NJ: Erlbaum.

Mobley, W. H. (1977). Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of Applied Psychology, 62,* 237–240.

Mobley, W. H., Griffeth, R. W., Hand, H. H., & Meglino, B. M. (1979). Review and conceptual analysis of the employee turnover process. *Psychological Bulletin, 86,* 493–522.

Moos, R. (1976). *The human context: Environmental determinants of behavior.* New York: Wiley.

Mowday, R. T., Koberg, C. S., & McArthur, A. W. (1984). The psychology of the withdrawal process: A cross-validational test of Mobley's intermediate linkages model of turnover in two samples. *Academy of Management Journal, 27,* 79–94.

Muchinsky, P. M., & Tuttle, M. L. (1979). Employee turnover: An empirical and methodological assessment. *Journal of Vocational Behavior, 14,* 43–77.

Muckler, F. A. (1982). Evaluating productivity. In M. D. Dunnette & E. A. Fleishman (Eds.), *Human performance and productivity: Vol. 1. Human capability assessment* (pp. 13–47). Hillsdale, NJ: Erlbaum.

Münsterberg, H. (1913). *Psychology and industrial efficiency.* Boston: Houghton-Mifflin.

Murphy, K. R. (1989a). Dimensions of job performance. In R. F. Dillon & J. W. Pellegrino (Eds.), *Testing: Theoretical and applied perspectives* (pp. 218–247). New York: Praeger.

Murphy, K. R. (1989b). Is the relationship between cognitive ability and job performance stable over time? *Human Performance, 2,* 183–200.

Murphy, K. R. (1991). Criterion issues in performance appraisal research: Behavioral accuracy versus classification accuracy. *Organizational Behavior and Human Decision Processes, 50,* 45–50.

Murphy, K. R., & Balzer, W. (1989). Rating errors and rater accuracy. *Journal of Applied Psychology, 74,* 619–624.

Murphy, K. R., & Cleveland, J. N. (1991). *Performance appraisal: An organizational perspective.* Boston: Allyn & Bacon.

Murphy, K. R., Herr, B. M., Lockhart, M. C., & Maguire, E. (1986). Evaluating the performance of paper people. *Journal of Applied Psychology, 71,* 654–661.

Murphy, K. R., & Kroeker, L. P. (1988). *Dimensions of job performance* (NPRDC TN 88-39). San Diego, CA: Navy Personnel Research and Development Center.

Nagle, B. (1953). Criterion development. *Personnel Psychology, 6,* 271–289.

Nathan, B. R., & Alexander, R. A. (1988). A comparison of criteria for test validation: A meta-analytic investigation. *Personnel Psychology, 41,* 517–535.

National Industrial Conference Board. (1938). *Plans for rating employees* (Studies in Personnel Policy No. 8). New York: Author.

Naylor, J. C., Pritchard, R. D., & Ilgen, D. R. (1980). *A theory of behavior in organizations.* New York: Academic Press.

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

Nesselroade, J. (1991). Interindividual differences in intraindividual change. In L. Collins & J. Horn (Eds.), *Best methods for the analysis of change* (pp. 92–105). Washington, DC: American Psychological Association.

Nesselroade, J., & Cattell, R. B. (Eds.). (1988). *Handbook of multivariate experimental psychology* (2nd ed.). New York: Plenum Press.

Nestor, O. W. (1986). *A history of personnel administration, 1890–1910.* New York: Garland.

Newbold, E. M. (1926). *A contribution to the study of the human factor in the causation of accidents* (Industrial Fatigue Research Board Report No. 34). London: His Majesty's Stationers Office.

Newcomb, T. R. (1931). An experiment designed to test the validity of a rating technique. *Journal of Educational Psychology, 22,* 279–289.

Novick, M. R. (1981). Federal guidelines and professional standards. *American Psychologist, 36,* 1035–1046.

Novick, M. R. (1982). Ability testing: Federal guidelines and professional standards. In A. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (pp. 70–98). Washington, DC: National Academy Press.

Office of Strategic Services. (1948). *Assessment of men.* New York: Rinehart.

Oppler, S. H., Campbell, J. P., Pulakos, E. D., & Borman, W. C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: Review, results, and conclusions. *Journal of Applied Psychology, 77,* 201–217.

Organ, D. (1988). *Organizational citizenship behavior: The good soldier syndrome.* Lexington, MA: Heath.

Otis, J. L. (1938). The prediction of success in power sewing machine operating. *Journal of Applied Psychology, 22,* 350–366.

Otis, J. L. (1940). The criterion. In W. H. Stead, C. L. Shartle, & Associates (Eds.), *Occupational counseling techniques* (pp. 73–94). New York: American Book.

Paese, P., & Switzer, F. S., III. (1988). Validity generalization and hypothetical reliability distributions: A test of the Schmidt-Hunter procedure. *Journal of Applied Psychology, 73,* 273–274.

Parker, G. V. C. (1971). Prediction of individual stability. *Educational and Psychological Measurement, 31,* 875–886.

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

Parsons, F. (1909). *Choosing a vocation.* Boston: Houghton-Mifflin.

Paterson, D. G. (1923a). Methods of rating human qualities. *Annals of the American Academy of Political and Social Science, 110,* 81–93.

Paterson, D. G. (1923b). The Scott Company graphic rating scale. *Journal of Personnel Research, 1,* 361–376.

Paterson, D. G. (1935). A target for critics. *Occupations, 13,* 18–21.

Patterson, C. H. (1946). On the problem of the criterion in predictive studies. *Journal of Consulting Psychology, 10,* 277–280.

Pearson, K. (1957). *The grammar of science.* New York: Meridian Books. (Original work published 1892).

Peel, M. D., & Alexander, C. (1936). Bibliography. *Occupations, 14,* 968–975.

Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influence of a frequently overlooked construct. *Academy of Management Review, 5,* 391–397.

Peters, L. H., O'Connor, E. J., Eulberg, J. R. (1985). Situational constraints: Sources, consequences, and future considerations. In K. M. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 3, pp. 79–113). Greenwich, CT: JAI Press.

Peters, L., & Sheridan, J. E. (1988). Turnover research methodology: A critique of traditional designs and a suggested survival analysis alternative. In K. M. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 6, pp. 231–262). Greenwich, CT: JAL Press.

Pond, M. (1927). Selective placement of metalworkers. *Journal of Personnel Research, 5,* 345–368, 405–417, 452–466.

Porter, L. W., & Lawler, E. E., III (1968). *Managerial attitudes and performance.* Homewood, IL: Irwin.

Porter, L. W., & Steers, R. M. (1973). Organizational, work and personal factors in employee turnover and absenteeism. *Psychological Bulletin, 80,* 151–176.

Price, J. L. (1977). *The study of turnover.* Ames: Iowa State University Press.

Price Waterhouse v. Hopkins, 109 S. Ct. 1775 (1989).

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69,* 581–588.

Pulakos, E. D., Schmitt, N., & Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within raters to measure halo. *Journal of Applied Psychology, 71,* 29–32.

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology, 74,* 770–780.

Rambo, W. W., Chomiak, A., & Price, J. M. (1983). Consistency of performance under stable conditions of work. *Journal of Applied Psychology, 68,* 78–87.

Reber, A. W. (1985). *Dictionary of psychology.* London: Penguin.

Richards, T. W., & Simons, M. P. (1941). The Fels child behavior scales. *Genetic Psychology Monographs, 24,* 259–309.

Richardson, M. W., & Kuder, G. F. (1933). Making a rating scale that measures. *Personnel Journal, 12,* 36–40.

Roach, D. E., & Wherry, R. J., Sr. (1970). Performance dimensions of multi-line insurance agents. *Personnel Psychology, 23,* 239–250.

Roback, A. A. (1917). The moral issues involved in applied psychology. *Journal of Applied Psychology, 1,* 232–243.

Robinson, E. S. (1919). The analysis of trade ability. *Journal of Applied Psychology, 3,* 352–357.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92,* 726–748.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika, 50,* 203–228.

Ronan, W. W., & Prien, E. (1966). *Toward a criterion theory: A review and analysis of research and opinion.* Greensboro, NC: Smith Richardson Foundation.

Ronan, W. W., & Prien, E. (1971). *Perspectives on the measurement of human performance.* New York: Appleton-Century-Crofts.

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

Rothe, H. (1978). Output rates among industrial employees. *Journal of Applied Psychology, 63,* 40–46.

Rousseau, D. M. (1985). Issues of level in organizational research: Multilevel and cross-level perspectives. In L. L. Cummings & B. M Staw (Eds.), *Research in organizational behavior* (Vol. 7, pp. 1–38). Greenwich, CT: JAI Press.

Rowe, P. M. (1967). Order effects in assessment decisions. *Journal of Applied Psychology, 51,* 170–173.

Roznowski, M., & Hanisch, K. A. (1989). Building systematic heterogeneity into measures of work attitudes and behavior measures. *Journal of Vocational Behavior, 36,* 361–375.

Rugg, H. (1921). Is the rating of human character practicable? *Journal of Educational Psychology, 12,* 425–438, 485–501.

Rugg, H. (1922). Is the rating of human character practicable? *Journal of Educational Psychology, 13,* 30–42, 81–93.

Rush, C. H., Jr. (1953). A factorial study of sales criteria. *Personnel Psychology, 6,* 9–24.

Rush, M. C., Phillips, J. S., & Lord, R. G. (1981). The effects of a temporal delay on leader behavior descriptions: A laboratory investigation. *Journal of Applied Psychology, 66,* 442–450.

Ryans, D. G. (1960). *Characteristics of teachers.* Washington, DC: American Council on Education.

Ryans, D. G., & Frederiksen, N. (1951). Performance tests of educational achievement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 455–494). New York: American Council on Education.

Saal, F., Downey, R., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88,* 413–428.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum performance. *Journal of Applied Psychology, 73,* 482–486.

Salmon, W. C. (1967). *The foundations of scientific inference.* Pittsburgh, PA: University of Pittsburgh Press.

Salmon, W. C., Jeffrey, R. C., & Greeno, J. G. (1971). *Statistical explanation and statistical relevance.* Pittsburgh, PA: University of Pittsburgh Press.

Samelson, F. (1977). World War I intelligence testing and the development of psychology. *Journal of the History of the Behavioral Sciences, 13,* 274–282.

Sauser, W. I., & Pond, S. B. (1981). Effects of rater training and participation on cognitive complexity: An exploration of Schneier's cognitive reinterpretation. *Personnel Psychology, 34,* 609–626.

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

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71,* 432–439.

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

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10,* 1–22.

Schmitt, N. W. (1989). Construct validity in personnel selection. In B. J. Fallon, H. Pfister, & J. Brebner (Eds.), *Advances in industrial and organizational psychology* (pp. 331–341). Amsterdam: North-Holland.

Schmitt, N. W., & Klimoski, R. J. (1991). *Research methods in human resources management.* Cincinnati, OH: South-Western.

Schneider, B. (1983). Interactional psychology and organizational behavior. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior* (Vol. 5, pp. 1–31). Greenwich, CT: JAI Press.

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

Schneier, C. E., & Beatty, R. W. (1979). *Performance appraisal sourcebook.* Amherst, MA: Human Resource Development Press.

Schoorman, F. D., & Schneider, B. J. (Eds.). (1988). *Facilitating work effectiveness.* Lexington, MA: Lexington.

Schultz, D. G., & Siegel, A. I. (1963). Progress and problems in the measurement of individual differences in on-the-job performance. *Acta Psychologica, 21,* 120–156.

Schwab, D. T. (1980). Construct validity and organizational behavior. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 2, pp. 3–43). Greenwich, CT: JAI Press.

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

Scott, W. D. (1917). A fourth method of checking results in vocational selection. *Journal of Applied Psychology, 1,* 61–67.

Scott, W. D. (1920). Changes in some of our conceptions and practices of personnel. *Psychological Review, 27,* 81–94.

Scott, W. D., & Clothier, R. C. (1923). *Personnel management.* New York: McGraw-Hill.

Scott, W. E., Jr., & Hamner, W. C. (1975). 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, 14,* 360–370.

Seashore, R. H. (1939). Work methods: A often neglected factor underlying individual differences. *Psychological Review, 46,* 123–141.

Seashore, S. E., Indik, B. P., & Georgopoulos, B. S. (1960). Relationships among criteria of job performance. *Journal of Applied Psychology, 44,* 195–202.

Sells, S. B. (Ed.). (1963). *Stimulus determinant of behavior.* New York: Ronald Press.

Sells, S. B. (1966). Multivariate technology in industrial and military personnel psychology. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 841–855). Chicago: Rand McNally.

Severin, D. (1952). The predictability of various kinds of criteria. *Personnel Psychology, 5,* 93–104.

Shaffer, L. F. (1936). *The psychology of adjustment.* Boston: Houghton-Mifflin.

Sheehy, N. P., & Chapman, A. J. (1987). Industrial accidents. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology 1987* (Vol. 2, pp. 201–227). Chichester, England: Wiley.

Shellow, S. M. (1925–1926). Selection of motormen in Milwaukee. *Journal of Personnel Research, 4,* 222–237.

Shen, E. (1925). The validity of self-estimate. *Journal of Educational Psychology, 16,* 104–107.

Shen, E. (1926). The reliability coefficient of personal ratings. *Journal of Educational Psychology, 16,* 232–236.

Sisson, E. D. (1948). Forced choice: The new army rating. *Personnel Psychology, 1,* 365–381.

Slavin R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15*(9), 5–11.

Slovic, P., & Lichtenstein, S. C. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6,* 649–744.

Smith, M., & Robertson, I. T. (Eds.). (1989). *Advances in selection and assessment.* Chichester, England: Wiley.

Smith, P. C. (1976). Behavior, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 745–775). Chicago: Rand McNally.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations. *Journal of Applied Psychology, 47,* 149–155.

Snow, A. J. (1926). Tests for chauffeurs. *Industrial Psychology, 1,* 30–45.

Sokal, M. M. (1984). James McKeen Cattell and American psychology of the 1920s. In J. Brozek (Ed.), *Explorations in the history of psychology in the United States* (pp. 273–323). Lewisburg, PA: Bucknell University Press.

Sokal, M. M. (Ed.). (1987). *Psychological testing and American society, 1890–1930.* New Brunswick, NJ: Rutgers University Press.

Spool, M. D. (1978). Training programs for observers of behavior: A review. *Personnel Psychology, 31,* 853–888.

Staff, Personnel Research Section. (1946). The forced choice technique and rating scales. *American Psychologist, 1,* 267.

Starr, R. B., & Greenly, R. (1938–1939). Merit rating survey findings. *Personnel Journal, 17,* 378–384.

Stead, W., Shartle, C. L., & Associates. (1940). *Occupational counseling techniques.* New York: American Book.

Steel, R. P., & Mento, A. J. (1986). Impact of situational constraints on subjective and objective criteria of managerial job performance. *Organizational Behavior and Human Decision Progress, 37,* 254–265.

Steers, R. M., & Porter, L. W. (1974). The role of task-goal attributes in employee performance. *Psychological Bulletin, 81,* 434–451.

Steers, R. M., & Rhodes, S. (1978). Major influences on employee attendance: A process model. *Journal of Applied Psychology, 63,* 391–407.

Steffy, B. D., & Maurer, S. D. (1988). Conceptualizing and measuring the economic effectiveness of human resource activities. *Academy of Management Review, 13,* 271–296.

Stern, G. G., Stein, M. I., & Bloom, B. S. (1956). *Methods in personality assessment.* Glencoe, IL: Free Press.

Stott, M. B. (1936). Criteria used in Britain. *Occupations, 14,* 953–957.

Stott, M. B. (1939). Occupational success. *Occupational Psychology, 13,* 126–140.

Strong, E. K., Jr. (1918). Work of the committee on classification of personnel in the army. *Journal of Applied Psychology, 2,* 130–139.

Strong, E. K., Jr. (1955). Walter Dill Scott: 1869–1955. *American Journal of Psychology, 68,* 682–683.

Strong, E. K., Jr., & Uhrbrock, R. S. (1923). *Job analysis and the curriculum.* Baltimore: Warwick.

Stuit, D. B. (Ed.). (1947). *Personnel research and test development in the*

Bureau of Naval Personnel. Princeton, NJ: Princeton University Press.

Sulsky, L. W., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73,* 497–506.

Surber, C. (1984). Issues in using quantitative rating scales in developmental research. *Psychological Bulletin, 95,* 226–246.

Symonds, P. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology, 7,* 456–461.

Symonds, P. (1925). Notes on rating. *Journal of Applied Psychology, 9,* 188–195.

Symonds, P. (1931). *Diagnosing personality and conduct.* New York: Century.

Taft, R. (1955). The ability to judge people. *Psychological Bulletin, 52,* 1–23.

Taylor, E. K., Barrett, R. S., Parker, J. W., & Martens, L. (1958). Rating scale content: II. Effect of rating on individual scales. *Personnel Psychology, 11,* 519–533.

Taylor, E. K., Parker, J. W., & Ford, G. L. (1959). Rating scale content: IV. Predictability of structured and unstructured scales. *Personnel Psychology, 12,* 247–266.

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

Taylor, F. W. (1947). *Principles of scientific management.* New York: Norton. (Original work published 1911).

Thorndike, E. L. (1918). Fundamental theorems in judging men. *Journal of Applied Psychology, 2,* 67–76.

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

Thorndike, E. L. (1935). Rebounds from the target. *Occupations, 13,* 329–333.

Thorndike, E. L., Bregman, E. O., Lorge, I., Metcalfe, Z. F., Robinson, E. E., & Woodyard, E. (1934). *Prediction of vocational success.* New York: Commonwealth Fund.

Thorndike, R. L. (1947). *Research problems and techniques* (Army-Air Force Aviation Psychology Program Research Rep. No. 3). Washington, DC: U.S. Government Printing Office.

Thorndike, R. L. (1949). *Personnel selection.* New York: Wiley.

Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). New York: American Council on Education.

Thorndike, R. L. (1991). Edward L. Thorndike: A professional and personal appreciation. In G. A. Kimble, M. Wertheimer, & C. L. White (Eds.), *Portraits of pioneers in psychology* (pp. 139–151). Hillsdale, NJ: Erlbaum.

Thorndike, R. L., & Hagen, E. (1959). *10,000 careers.* New York: Wiley.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review, 38,* 406–427.

Tiffin, J. (1942). *Industrial psychology.* New York: Prentice-Hall.

Toops, H. A. (1921). *Trade tests in education.* New York: Teachers College, Columbia University.

Toops, H. A. (1944). The criterion. *Educational and Psychological Measurement, 4,* 271–297.

Toops, H. A. (1959). A research utopia in industrial psychology. *Personnel Psychology, 12,* 189–225.

Travers, R. M. W. (1951). A critical review of the validity and rationale of the forced-choice technique. *Psychological Bulletin, 48,* 62–70.

Tziner, A., & Kopelman, R. (1988). Effects of rating format on goal-setting dimensions. *Journal of Applied Psychology, 73,* 323–328.

Uhlaner, J. E. (1972). Human performance effectiveness and the systems measurement bed. *Journal of Applied Psychology, 56,* 202–210.

Uhlaner, J. E., & Drucker, A. J. (1964). Criteria for human performance research. *Human Factors, 8,* 265–278.

Uhlaner, J. E., & Drucker, A. J. (1980). Military research on performance criteria: A change of emphasis. *Human Factors, 22,* 131–139.

Uhrbrock, R. S. (1950). Standardization of 724 rating scale statements. *Personnel Psychology, 3,* 285–316.

Uhrbrock, R. S. (1961). 2000 scaled items. *Personnel Psychology, 14,* 375–420.

Uniform Guidelines on Employee Selection Procedures, 43 Fed. Reg., 38290-38315 (1978).

Vallance, T. R., Glickman, A. S., & Suci, G. J. (1953). Criterion rationale for a personnel research program. *Journal of Applied Psychology, 37,* 429–431.

Vance, R. J., Coovert, M. D., MacCallum, R. C., & Hedge, J. W. (1989). Construct models of task performance. *Journal of Applied Psychology, 74,* 447–455.

Vance, R. J., MacCallum, R. C., Coovert, M. D., & Hedge, J. W. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. *Journal of Applied Psychology, 73,* 74–80.

Van Dusen, A. C. (1947). Importance of criteria in selection and training. *Educational and Psychological Measurement, 7,* 498–504.

Vernon, P. E. (1938). *The assessment of psychological qualities by verbal methods* (Industrial Health Research Board, Report No. 83). London: His Majesty's Stationery Office.

Villanova, P. (1992). A customer-based model for developing job performance criteria. *Human Resource Management Review, 2,* 103–114.

Villanova, P., & Bernardin, H. J. (1989). Impression management in the context of performance appraisal. In R. A. Giacalone & P. Rosenfeld (Eds.), *Impression management in the organization* (pp. 299–313). Hillsdale, NJ: Erlbaum.

Vinchur, A. J., Schippmann, J. S., Smalley, M. D., & Rothe, H. F. (1991). Productivity consistency of foundry chippers and grinders: A 6-year field study. *Journal of Applied Psychology, 76,* 134–136.

Viteles, M. S. (1925). The clinical viewpoint in vocational psychology. *Journal of Applied Psychology, 9,* 131–138.

Viteles, M. S. (1925–1926a). Selection of motormen. *Journal of Personnel Research, 4,* 100–115, 173–199.

Viteles, M. S. (1925–1926b). Standards of accomplishment: Criteria of vocational selection. *Journal of Personnel Research, 4,* 483–486.

Viteles, M. S. (1931). Industry. In R. A. Brotemarkle (Ed.), *Clinical psychology* (pp. 117–133). Philadelphia: University of Pennsylvania Press.

Viteles, M. S. (1932). *Industrial psychology.* New York: Norton.

Viteles, M. S. (1936). A dynamic criterion. *Occupations, 14,* 963–967.

Viteles, M. S. (1945). The aircraft pilot: Five years or research; A summary of outcomes. *Psychological Bulletin, 42,* 489–526.

Viteles, M. S. (1974). Industrial psychology: Reminiscences of an academic moonlighter. In T. Krawiec (Ed.), *The psychologists* (Vol. 3, pp. 441–500). New York: Oxford University Press.

Vondracek, F. W., Lerner, R. M., & Schulenberg, J. E. (1983). The concept of development in vocational theory and intervention. *Journal of Vocational Behavior, 23,* 179–202.

von Mayrhauser, R. (1987). The manager, the medic, and the mediator: The clash of professional psychological styles and the wartime origins of group mental testing. In M. M. Sokal (Ed.), *Psychological testing and American society, 1890–1930* (pp. 128–157). New Brunswick, NJ: Rutgers University Press.

von Mayrhauser, R. (1992). The mental testing community and validity: A prehistory. *American Psychologist, 47,* 244–253.

Vroom, V. (1964). *Work and motivation.* New York: Wiley.

Waldman, D. A., & Avolio, B. J. (1986). A meta-analysis of age differences in job performance. *Journal of Applied Psychology, 71,* 33–38.

Wallace, S. R. (1965). Criteria for what? *American Psychologist, 20,* 411–417.

Wallace, S. R., & Weitz, J. (1955). Industrial psychology. *Annual Review of Psychology, 6,* 217–250.

Wards Cove Packing Co. v. Atonio, 109 S. Ct. 2115 (1988).

Warren, H. C. (Ed.). (1934). *Dictionary of psychology.* Boston: Houghton-Mifflin.

Watson v. Ft. Worth Bank and Trust, 108 S. Ct. 2777 (1988).

Wechsler, D. (1926–1927). Tests for taxicab drivers. *Journal of Personnel Research, 5,* 24–30.

Weiss, L. A. (1933). Rating scales. *Psychological Bulletin, 30,* 185–208.

Weitz, J. (1961). Criteria for criteria. *American Psychologist, 16,* 228–232.

Wells, F. L. (1907). A statistical study of literary merit. *Archives of Psychology, 7* (X).

Wendelken, D. J., & Inn, A. (1981). Nonperformance influences on performance evaluations: A laboratory phenomenon? *Journal of Applied Psychology, 66,* 149–158.

Wernimont, P., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372–376.

Wherry, R. J., Sr. (1950). Criteria and validity. In D. Fryer & E. R. Henry (Eds.), *Handbook of applied psychology* (Vol. 1, pp. 170–177). New York: Rinehart.

Wherry, R. J., Sr. (1952). *The control of bias in rating: A theory of rating* (Personnel Research Board Report 922). Washington, DC: Department of the Army, Adjutant General's Office, Personnel Research Section.

Wherry, R. J., Sr. (1957). The past and future of criterion evaluation. *Personnel Psychology, 10,* 1–5.

Wherry, R. J., Sr. (1959). Hierarchical factor solutions without rotations. *Psychometrika, 24,* 45–51.

Wherry, R. J., Sr., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35,* 521–551.

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

Whitlock, G. H. (1963). Applications of the psychophysical law to performance evaluation. *Journal of Applied Psychology, 47,* 15–23.

Wiebe, R. H. (1967). *The search for order.* New York: Hill & Wang.

Wigdor, A., & Garner, W. R. (Eds.). (1982). *Ability testing: Uses, consequences, and controversies.* Washington, DC: National Academy Press.

Wigdor, A., & Green, B. F., Jr. (Eds.). (1986). *Assessing the performance of enlisted personnel: Evaluation of a joint-service research project.* Washington, DC: National Academy Press.

Wigdor, A., & Green, B. F., Jr. (Eds.). (1991). *Performance assessment for the workplace* (2 vols.). Washington, DC: National Academy Press.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. Snow & D. E. Wiley (Eds.), *Improving inquiry in the social sciences* (pp. 75–107). Hillsdale, NJ: Erlbaum.

Wyer, R. S., & Srull, T. (1989). *Memory and cognition in its social context.* Hillsdale, NJ: Erlbaum.

Yerkes, R. M. (1917). The Binet version versus the point scale method of measuring intelligence. *Journal of Applied Psychology, 1,* 111–122.

Yoakum, C. S. (1922–1923). Basic experiments in vocational guidance. *Journal of Personnel Research, 1,* 18–34.

Yoakum, C. S., Manson, G. E. (1926). Self-ratings as a means of determining trait-relationships and relative desirability of traits. *Journal of Abnormal and Social Psychology, 21,* 52–64.

Yoakum, C. S., & Yerkes, R. S. (Eds.). (1920). *Army mental tests.* New York: Holt.

Zammuto, R. F. (1984). A comparison of multiple constituency models of organizational effectiveness. *Academy of Management Review, 9,* 606–616.

Zammuto, R. F., London, M., & Rowland. K. M. (1982). Organization and rater difference in performance appraisal. *Personnel Psychology, 35,* 643–658.

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

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

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

Zeidner, J. (Ed.). (1987). *Human productivity enhancement* (2 vols.). New York: Praeger.

Zerga, J. E. (1943). Developing an industrial merit rating scale. *Journal of Applied Psychology, 27,* 190–195.

Received August 12, 1991
Revision received June 15, 1992
Accepted June 15, 1992 ■

1993 APA Convention "Call for Programs"

The "Call for Programs" for the 1993 APA annual convention appears in the October issue of the APA Monitor. The 1993 convention will be held in Toronto, Ontario, Canada, from August 20 through August 24. Deadline for submission of program and presentation proposals is December 10, 1992. Additional copies of the "Call" are available from the APA Convention Office, effective in October. As a reminder, agreement to participate in the APA convention is now presumed to convey permission for the presentation to be audiotaped if selected for taping. Any speaker or participant who does not wish his or her presentation to be audiotaped must notify the person submitting the program either at the time the invitation is extended or prior to the December 10 deadline for proposal submission.