6

Assessment of Individual Job Performance: A Review of the Past Century and a Look Ahead

CHOCKALINGAM VISWESVARAN

Job performance is a central construct in work psychology. The methods of assessing individual job performance, the factor structure of the construct, criteria for evaluating the criterion, as well as path models explaining individual job performance, are reviewed. The factor structure of job performance is best conceptualized as a hierarchy, with the general factor at the apex and several group factors at the second, lower level. The number and nature of the group factors varies according to the theorist. Issues of bias in ratings as well as contamination and deficiency in nonratings measures are summarized. The evidence for the reliability and construct validity of individual job performance assessment (both for overall assessments as well as dimensional assessments) are presented. The changing nature of work and its impact on the conceptualization and assessment of individual job performance are noted.

Job performance is an important construct in industrial/organizational psychology (Arvey & Murphy, 1998; Austin & Villanova, 1992; Campbell, 1990; Murphy & Cleveland, 1995; Schmidt & Hunter, 1992). In fact, most of what industrial-organizational psychologists do is geared to have a positive impact on job performance. The importance of assessment of individual job performance is probably reflected in the volume of literature devoted to it, and many leading researchers in our field have written on the topic of individual job performance.

Individual job performance plays a central role in what we do as researchers and practitioners. For example, one question in recruitment is whether the different sources of recruitment result in attraction of individuals who differ in job performance levels (Barber, 1998). In personnel selection, attempts are made to identify individual differences variables that are related to individual differences in job performance, and select individuals based on those characteristics (Guion, 1998). Organizations require that the expenses associated with training programs (e.g., socialization or orientation programs, skills training) be justified with evidence that such training improves individual job performance. In short, individual job performance is a central construct in our field.

Individual job performance data can be used in numerous ways. Cleveland, Murphy and Williams (1989) identified several uses of individual job performance data, classifying these uses into four categories: (1) between-person decisions, (2) within-person decisions, (3) systems maintenance, and (4) documentation. Between-persons decisional uses included the use of individual job performance data for salary administration purposes, making promotion decisions, and to design merit pay systems. Within-person decisions included providing feedback to individuals so as to identify individual strengths and weaknesses – data that is used for

assessing training and placement needs. The systems maintenance category refers to the use of individual job performance assessments for human resources planning and reinforcement of authority structures in organizations. Finally, individual job performance data are also used for legal documentation purposes.

Cascio (1991) groups these uses into three main categories: administrative, feedback, and research purposes. Administrative use refers to the use of individual job performance assessment for making administrative decisions such as pay allocation, promotions, and layoffs. Individual job performance assessment can also be used to provide feedback to individuals by identifying their strengths and weaknesses, and finally, it is required for research purposes – be it validation of a selection technique or evaluating the efficacy of a training program.

That individual job performance assessments are used in a variety of ways has also been found in several studies. For a long time, administrative uses have been known (Whisler & Harper, 1962), and DeVries, Morrison, Shullman and Gerlach (1986) report that surveys conducted in the 1970s in both the United States and the United Kingdom indicated the prevalence of individual job performance assessment for the purpose of making administrative decisions. In fact, these surveys suggested that more than 50% of the use of individual job performance assessment was for the purpose of making administrative decisions. DeVries et al. (1986) noted that the use of such assessments in Great Britain can be classified into three categories: (1) to improve current performance, (2) to set objectives, and (3) to identify training and development needs.

In this chapter, I review the research on individual job performance. There are four sections. The first deals with the different methods of assessment, and following this I summarize the studies conducted to explicate the content domain of individual job performance. Factor analytic studies as well as theoretical and rational analyses of what constitutes individual job performance are reviewed. In the third section, I review the criteria for assessing the quality of individual job performance assessments along with a discussion of such studies. Finally, in the fourth section, I summarize some of the causal path models postulated to explain the determinants and components of individual job performance.

METHODS OF ASSESSMENT

Methods used to assess individual job performance can be broadly classified into (1) organizational records, and (2) subjective evaluations. Organizational records are considered to be more 'objective' in contrast to the subjective evaluations that depend on a human judgment. Subjective evaluations could either be criterion referenced (e.g., ratings) or norm-referenced (e.g., rankings). The distinction between organizational records and subjective evaluations has a long history. Burtt (1926) and Viteles (1932) grouped criterion measures into objective and subjective classes. Farmer (1933) grouped criteria into objective measures, judgments of performance (judgments based on objective performance), and judgments of ability (judgments based on traits). Smith (1976) distinguished between hard criteria (i.e., organizational records) and soft criteria (i.e., subjective evaluations).

Methods of assessments should be distinguished from types of criteria. Thorndike (1949) identifies three types of criteria: immediate, intermediate, and ultimate criteria. The ultimate criterion summarizes the total worth of the individual to the organization over the entire career span. The immediate criteria on the other hand is a measure of individual job performance at that particular point in time. Intermediate criteria summarize performance over a period of time. Note that both organizational records and subjective evaluations can be used to assess, say, an intermediate criterion. Similarly, Mace (1935) argued that measures of individual job performance can stress either capacity or will to perform. This distinction is a forerunner to the distinction between maximal and typical performance measures (e.g., DuBois, Sackett, Zedeck & Fogli, 1993; Sackett, Zedeck & Fogli, 1988). Maximal performance is what an individual can do if highly motivated whereas typical performance is what an individual is likely to do in a typical day. The distinction between ultimate, intermediate, and immediate criteria or between maximal and typical performance refers to types of criteria. Both organizational records and subjective evaluations (methods) can be used to assess them.

Organizational records can be further classified into direct measures of productivity and personnel data (Schmidt, 1980). Direct measures of productivity stress the number of units produced. Also included are measures of quality such as the number of errors, scrap material produced, and so forth. Personnel data, on the other hand, do not directly measure productivity but inferences of productivity can be derived based on them. Lateness or tardiness, tenure, absences, accidents, promotion rates, and filing grievances can be considered as indirect measures of productivity – there is an inferential leap involved in using these personnel data as a measure of individual job performance. Organizational records, by focusing on observable, countable, discrete outcomes, may overcome the biasing influences of subjective evaluations but may be affected by criterion contamination and criterion deficiency. Contamination occurs in that outcomes could be due to factors beyond the control of the individuals; deficiency results as the outcomes assessed may not take into account important aspects of individual

job performance. I will discuss the literature on the construct validity of organizational records after presenting the criteria for the job performance criterion in the third section of this chapter.

Subjective evaluations can be either ratings or rankings of performance. Ratings are criterion-referenced judgments where an individual is evaluated without reference to other individuals. The most common form of rating scale is a graphic rating scale (GRS), which typically involves presenting the rater with a set of dimensions or task categories with several levels of performance and requiring the raters to choose the level that best describes the person being rated. There are several formats of GRS. The different formats differ in the number of levels presented, the clarity in demarcating the different levels (e.g., asking the rater to circle a number vs. asking them to indicate a point in a line the end points of which are described), and the clarity in identifying what behaviors constitute a particular level. Smith and Kendall (1963) designed the Behaviorally Anchored Rating Scales (BARS) to explicitly tie the different levels to behavioral anchors. Steps involved in the construction of BARS include generating a list of behaviors depicting different performance levels of a particular dimension of performance, checking the agreement across raters (retranslation), and designing the layout of the scale. A variant of the BARS is the Behavioral Observation Scale (BOS) where the rater merely notes whether a behavior was displayed by the ratee (Latham Fay, & Saari, 1980) and the Behavioral Evaluation Scale (BES) where the rater notes the likelihood of the ratee exhibiting a particular behavior (Bernardin, Alvares & Cranny, 1976).

Researchers have also addressed, by developing checklists, the reluctance of raters to judge the performance of others. The rater merely indicates whether a particular behavior has been exhibited, and either a simple sum or weighted combination is then computed to assess performance. There are several types of these summated rating scales in existence. To address the problem that raters could intentionally distort their ratings, forced choice scales and mixed standard scales (MSS) have also been developed. In a forced choice assessment, raters are provided with two equally favorable statements of which only one discriminates between good and poor performers. The idea is that the rater who wants to give lenient ratings may choose the favorable but nondiscriminating statement as descriptive of the rater. The MSS comprises of three statements for each dimension of performance rated with the three statements depicting an excellent, an average and a poor performance, respectively on that dimension. The rater rates the performance of each ratee as better than, equal to or worse than the performance depicted in that statement. Scoring rules are developed and MSS can identify inconsistent or careless raters (Blanz & Ghiselli, 1972).

Several research studies have been conducted over the years to compare the quality of the different rating scales. Symonds (1924) investigated the optimal number of scale points and recommended seven categories as optimal. Other researchers (e.g., Bendig, 1954; Lissitz & Green, 1975) present conflicting conclusions. Schwab Heneman and DeCotiis (1975) questioned the superiority of BARS over other formats, and finally, Landy and Farr (1980) in an influential article concluded that rating formats and scales do not alter the performance assessments, and guided researchers away from the unprofitable controversies of which scale and rating format is superior to investigations of the cognitive processes underlying performance assessments.

In contrast to ratings which are criterion-referenced assessments, rankings are norm-referenced assessments. The simplest form of ranking is to rank all ratees from best to worst. The ranking will depend on the set of ratees and it is impossible to compare the rankings from two different sets of individuals; the worst in one set may be better than the best in the second set of ratees. A modified version, called alternate ranking, involves (1) picking the best and worst ratees in the set of ratees under consideration, (2) removing the two chosen ratees, (3) picking the next best and worst from the remaining ratees, and (4) repeating the process until all ratees are ranked. The advantage of the alternate ranking method is that it reduces the cognitive load on the raters. Yet another approach is to compare each ratee to every other ratee, a method of paired comparisons that becomes unwieldy when the number of ratees increases. Finally, forced distribution methods can be used where a fixed percentage of ratees are placed in each level. Forced distribution methods can be useful to generate the desired distribution (mostly normal) of assessed scores.

With subjective evaluations (ratings or rankings), the question of who should rate arises. Typically, in traditional organizations the supervisors of the employees provide the ratings. Recent years have seen an increase in the use of 360 degree feedback systems (Church & Bracken, 1997) where rating assessments can be made by self (the ratee himself or herself), subordinates, peers, and customers or clients. I discuss the convergence among the different sources as well as the convergence between subjective evaluations and organizational records under the section on the construct validity of performance assessments.

EXPLICATING THE CONSTRUCT DOMAIN OF INDIVIDUAL JOB PERFORMANCE

Job performance is an abstract, latent construct. One cannot point to one single physical manifestation and define it as job performance; there are several

manifestations of an abstract construct. Explicating the construct domain of individual job performance involves specifying what is included when we talk of the concept (Wallace, 1965). Further, keeping with the abstract nature of constructs, there are several manifestations of individual job performance with the actual operational measure varying across contexts; explication of the construct involves identifying dimensions that make up the construct. The dimensions generalize across contexts whereas the exact measures differ. For example, interpersonal competence is a dimension of individual job performance that could be relevant in several contexts, but the actual behavior could vary depending on the construct. One measure of interpersonal competence for a professor may be how polite the professor is in replying to reviewers. For a bank teller, a measure of interpersonal competence is how considerate they are of customer complaints or the extent to which they smile at customers.

To explicate a construct domain, it is optimal to start with a definition of the construct. In this chapter, I define individual job performance as evaluatable behaviors. Although I use the term behaviors, I would stress that the difference between behaviors and outcomes is not clear-cut in many instances. Some researchers (Campbell, 1990) insist on a clear demarcation between behaviors and outcomes whereas others (Austin & Villanova, 1992; Bernardin & Pence, 1980) deemphasize this difference. The reason for emphasizing this difference between behaviors and outcomes is the alleged control an individual has over them. The argument is that the construct of individual job performance should not include what is beyond the individual's control. The distinguishing feature is whether the individual has control over what is assessed. If the individual does have such control, it is included under the individual job performance construct.

Consider the research productivity of a professor. Is the number of papers *published* a measure of individual job performance? Surely, several factors beyond the control of the professor affect the publishing of the paper. Is the number of papers *written*, a measure of individual job performance? Again, surely we can think of several factors that could affect the number of papers written that are not under the control of the professor. Thus, for every measure or index of individual job performance, the degree of control the individual has is a matter of degree. As such the distinction between behaviors and outcomes is also a question of degree and not some absolute distinction. Whether one defines performance and related constructs as behaviors or outcomes depends on the attributions one makes and the purpose of the evaluation.

How have researchers and practitioners defined the construct domain of individual job performance in their studies? Generally they have applied some combination of the following three approaches. First, researchers have reviewed job performance measures used in different contexts and attempted to synthesize what dimensions make up the construct. This rational method of synthesizing and theory building is however affected by the personal bias of the individual researchers.

Second, researchers have developed measures of hypothesized dimensions, collected data on these measures, and factor analyzed the data (e.g., Rush, 1953). This empirical approach is limited by the number and type of measures included in the data collection phase. Recently, Viswesvaran (1993) invoked the lexical hypothesis from personality literature (Goldberg, 1995) to address this limitation. The lexical hypothesis states that practically significant individual differences in personality are encoded in the language used, and therefore a comprehensive description of personality can be obtained by collating all the adjectives found in the dictionary. Viswesvaran, Ones and Schmidt (1996) extended this principle to job performance assessment and argued that a comprehensive specification of the content domain of the job performance construct can be obtained by collating all the measures of job performance that had been used in the extant literature.

Third, researchers (e.g., Welbourne, Johnson & Erez, 1998) have invoked organizational theories to define what the content of the job performance construct should be. Welbourne et al., used role theory and identity theory to explicate the construct of job performance. Another example of invoking a theory of work organization to explicate the construct of job performance comes in the distinction made between task and contextual performance (Borman & Motowidlo, 1993). Distinguishing between task and contextual performance parallels the social and technical systems that are postulated to make-up the organization. Of these three approaches, most of the extant literature employs either rational synthesis or factor analytic approaches. Therefore, I review these two set of studies separately.

Rational Synthesis of Job Performance Dimensions

Toops (1944) was one of the earliest attempts to hypothesize what dimensions comprise the construct of job performance, arguing a distinction between accuracy (quality or lack of errors) and volume of output (quantity). Toops (1944) lists units of production, quality of work, tenure, supervisory and leadership abilities as dimensions of individual job performance. Wherry (1957), on the other hand, lists listed six dimensions: output, quality, lost time, turnover, training time or promotability, and satisfaction. The last two decades have seen several rational analyses (of the individual job

performance construct) based on the plethora of factor analytic studies that have been conducted over the years. In this section, I present three such frameworks.

Bernardin and Beatty (1984) define performance as the record of outcomes produced on a specified job function or activity during a specified time period. Although a person's job performance depends on some combination of ability, motivation and situational constraints, it can be measured only in terms of some outcomes. Bernardin and Beatty (1984) then consider the issue of dimensions of job performance. Every job function could be assessed in terms of six dimensions (Kane, 1986): quality, quantity, timeliness, cost-effectiveness, need for supervision, and interpersonal impact. Some of these dimensions may not be relevant to all job activities. Bernardin and Russell (1998) emphasize the need to understand the interrelationships among the six dimensions of performance. For example, a work activity performed in sufficient quantity and quality but not in time may not be useful to the organization.

Campbell (1990) describes the latent structure of job performance in terms of eight dimensions. According to Campbell (1990) and Campbell, McCloy, Oppler and Sager (1993), the true score correlations between these eight dimensions are small, and hence any attempt to cumulate scores across the eight dimensions will be counterproductive for guiding research and interpreting results. The eight factors are: job-specific task proficiency, nonjob-specific task proficiency, written and oral communication, demonstrating effort, maintaining personal discipline, facilitating peer and team performance, supervision, and management or administration.

Job-specific task proficiency is defined as the degree to which the individual can perform the core substantive or technical tasks that are central to a job and which distinguish one job from another. Nonjob-specific task proficiency, on the other hand, is used to refer to tasks not specific to a particular job, but is expected of all members of the organization. Demonstrating effort captures the consistency or perseverance and intensity of the individuals to complete the task, whereas maintenance of personal discipline refers to the eschewment of negative behaviors (such as rule infractions) at work. Management or administration differs from supervision in that the former includes performance behaviors directed at managing the organization that are distinct from supervisory or leadership roles. Written and oral communications reflect that component of the job performance that refers to the proficiency of an incumbent to communicate (written or oral) independent of the correctness of the subject matter. The description of these eight dimensions are further elaborated in Campbell (1990) and Campbell et al. (1993). Five of the eight dimensions were found in a sample of military jobs (Campbell, McHenry & Wise, 1990). Further details about these dimensions may be found in Campbell (1990).

Murphy (1989) describes the construct of job performance as comprising of four dimensions: downtime behaviors, task performance, interpersonal, and destructive behaviors. Task performance focuses on performing role-prescribed activities whereas downtime behaviors refer to lateness, tardiness, absences or, broadly, to the negative pole of time on task (i.e., effort exerted by an individual on the job). Interpersonal behaviors refer to helping others, teamwork ratings, and prosocial behaviors. Finally, destructive behaviors correspond to compliance with rules (or lack of it), violence on the job, theft, and other behaviors counterproductive to the goals of the organization. The four dimensions are further elaborated in Murphy (1989).

Factor Analytic Studies

In a typical factor analytic study, individuals are assessed on multiple measures of job performance. Correlations are obtained between the measures of job performance and factor analysis is used to identify the measures that cluster together. Based on the commonalities across the measures that cluster together, a dimension is defined. For example, when absence measures, lateness measures, and tenure cluster together, a dimension of withdrawal is hypothesized. I review below some representative studies; the actual number of studies is too numerous to even list, let alone describe in a book chapter.

An important point needs to be stressed here. Factor analyses of importance ratings of task elements, frequency of tasks performed, and time spent on tasks done on the job, are not reviewed. The dimensions identified in such studies do not capture dimensions of individual job performance (Schmidt, 1980; Viswesvaran, 1993). Consider a typical job analytic study that obtains importance ratings of task statements from raters. The correlation between these ratings (e.g., the correlation between task *i* and task *j*) are computed and the resulting correlation matrix is factor analyzed. Tasks that cluster together are used to identify a dimension of job performance. But because all raters are rating the same stimulus (say task *i*), the true variance is zero (Schmidt & Hunter, 1989). Any observed variability across raters is the result of random errors, disagreements between raters, and differences between raters in leniency and other rater idiosyncrasies. Correlating the rating errors in pairs of variables (importance ratings of tasks *i* and *j*) and factor analyzing the resulting correlations cannot reveal individual differences dimensions of job performance (Schmidt, personal communication, June 25, 1993). Therefore, in this section I focus only on studies that obtained individual job performance

data on different measures, correlated the measures, and factor analyzed the resulting correlation matrix to identify dimensions of performance.

Rush (1953) factor analyzed nine rating measures and three organizational records-based measures of job performance for 100 salespeople. He identified the following four factors: objective achievement, learning aptitude, general reputation, and proficiency of sales techniques. A sample size of 100 for analyzing a 12 × 12 matrix of correlations would probably be considered inadequate by present-day standards, but this was one of the first studies to employ factor analytic techniques to explicate the underlying dimensions and factor structure of the individual job performance construct.

Baier and Dugan (1957) obtained data on 346 sales agents on 15 objective variables and two subjective ratings. Factor analysis of the 17 × 17 intercorrelation matrix resulted in one general factor. Several different measures such as percentage sales, units sold, tenure, knowledge of products, loaded on this general factor. In contrast, Prien and Kult (1968) factor analyzed a set of 23 job performance measures and found evidence for seven distinct dimensions. Roach and Wherry (1970) using a large sample of (N = 900) salespersons found evidence for a general factor. Seashore, Indik and Georgopoulos (1960) using comparably large samples (N = 975) found no evidence for a general factor.

Ronan (1963) conducted a factor analysis of a set of 11 job performance measures. Four of the measures were objective records including measures of accidents and disciplinary actions. The factor analysis indicated a four-factor solution. One of the four factors reflected the 'safe' work habits of the individual (e.g., index of injuries, time lost due to accidents); acceptance of authority and adjustment constituted two other factors. The fourth factor was uninterpretable (Ronan, 1963).

Gunderson and Ryman (1971) examined the factor structure of individual job performance in extremely isolated groups. The sample analyzed involved scientists spending their winter in Antarctica. Three factors were identified: task efficiency, emotional stability, and interpersonal relations. Klimoski and London (1974) obtained data from different sources (e.g., supervisors, peers) to avoid monomethod problems, and reported evidence for the presence of a general factor, a finding that is interesting when considered in the wake of arguments (cf. Borman, 1974) that raters at different levels of job performance construe the content domain of job performance differently.

Factor analytic studies in the last two decades (1980–99) have used much larger samples and refined techniques of factor analysis, and the use of confirmatory factor analysis has enabled researchers to combine rational synthesis and empirical partitioning of variance. For example, Borman, Motowidlo, Rose and Hansen (1985) developed a model of soldier effectiveness based on data collected during Project A. Project A is a multi-year effort undertaken by the United States Army to develop a comprehensive model of work effectiveness. As part of that landmark project, Borman et al., developed a model of job performance for first-tour soldiers that are important for unit effectiveness. Borman et al., noted that in addition to task performance, there were three performance dimensions: allegiance, teamwork, and determination, and that each of these three dimensions could be further subdivided. Thus, allegiance involved following orders, following regulations, respect for authority, military bearing, and commitment. Teamwork comprised of cooperation, camaraderie, concern for unit morale, boosting unit morale, and leadership. Determination involved perseverance, endurance, conscientiousness, initiative, and discipline.

Hunt (1996) developed a model of generic work behavior applicable to entry-level jobs especially in the service industry. Using performance data from over 18,000 employees primarily from the retail sector, Hunt identified nine dimensions of job performance that do not depend on job-specific knowledge. The nine dimensions were: adherence to confrontational rules, industriousness, thoroughness, schedule flexibility, attendance, off-task behavior, unruliness, theft, and drug misuse. Adherence to confrontational rules reflected an employee's willingness to follow rules that might result in a confrontation between the employee and a customer (e.g., checking for shoplifting). Industriousness captured the constant effort and attention towards work while on the job. Thoroughness was related to the quality of work whereas schedule flexibility reflected the employees' willingness to change their schedule to accommodate demands at work. Attendance captured the employee's presence at work when scheduled to work, and punctuality. Off-task behavior involved the use of company time to engage in nonjob activities. Unruliness referred to minor deviant tendencies as well as abrasive and inflammatory attitudes towards coworkers, supervisors, and work itself. Finally, theft involved taking money or company property, or helping friends steal property whereas drug misuse referred to inappropriate use of drugs and alcohol.

Another trend discernible in the last two decades is the focus on specific performance aspects other than task performance. Smith, Organ and Near (1983) popularized the concept of 'Organizational Citizenship Behavior' (OCB) into the job performance literature. OCB was defined as individual behavior that is discretionary, not directly or explicitly recognized by the formal reward system, and that in the aggregate promotes the effective functioning of the organization (Organ, 1988). Factor analytic studies have identified distinct sub-dimensions of OCB: altruism, courtesy, cheerleading, sportsmanship, civic virtue, and conscientiousness.

Over the years several concepts related and overlapping with OCB have been proposed. George and Brief (1992) introduced the concept of 'organizational spontaneity', defining organizational spontaneity as voluntarily performed extra-role behavior that contributes to organizational effectiveness. Five dimensions were postulated: helping co-workers, protecting the organization, making constructive suggestions, developing oneself, and spreading goodwill. Organizational spontaneity is distinguished from OCB partly on account of reward systems being designed to recognize organizational spontaneity.

Van Dyne, Cummings and Parks (1995) argued for the use of 'Extra-Role Behavior' (ERB). Based on role theory concepts developed by Katz (1964), ERB has been hypothesized to contribute to organizational effectiveness. Brief and Motowidlo (1986) introduced the related concept of Prosocial Organizational Behavior (POB), which has been defined as behavior performed with the intention of promoting the welfare of individuals or groups to whom the behavior has been directed. POB can be either role-prescribed or extra-role, and it can be negative towards organizations although positive towards individuals.

Finally, Borman (1991) as well as Borman and Motowidlo (1993) describe the construct of job performance as comprising task and contextual performance. Briefly, task performance focuses on performing role-prescribed activities whereas contextual performance accounts for all other helping and productive behaviors (Borman, 1991; Borman & Motowidlo, 1993). The two dimensions are further elaborated in Borman and Motowidlo (1993). Motowidlo, Borman and Schmit (1997) developed a theory of individual differences in task and contextual performance. Some researchers (e.g., Van Scotter & Motowidlo, 1996) have argued that individual differences in personality variables are linked more strongly than individual differences in (cognitive) abilities to individual differences in contextual performance. Cognitive ability was hypothesized to be more predictive of task performance than contextual performance. Although persuasive, empirical support for this argument has been mixed. Conscientiousness, a personality variable, has been linked as strongly as cognitive ability to task performance in some studies (Alonso, 2000).

Behaviors that have negative value for organizational effectiveness have also been proposed as constituting distinct dimensions of job performance, and organizational misbehavior has become a topic of research interest. Clark and Hollinger (1983) discussed the antecedents of employee theft on organizations. Our work on integrity testing (Ones, Viswesvaran & Schmidt, 1993) as well as the works of Paul Sackett and colleagues (cf. Sackett & Wanek, 1996) have identified the different forms of counterproductive behaviors such as property damage, substance abuse, violence on the job. Withdrawal behaviors have long been studied by work psychologists in terms of lateness or tardiness, absenteeism, and turnover. Work psychologists and social psychologists have explored the antecedents and consequences of social loafing, shirking or the propensity to withhold effort (Kidwell & Bennett, 1993).

A major concern in evaluating the different factor analytic studies in the job performance domain is the fact that the dimensions identified are a function of the measures included. To ensure a comprehensive specification of the content domain of the job performance construct, Viswesvaran (1993) invoked the lexical hypothesis which was first introduced in the personality assessment literature (see also Viswesvaran et al., 1996). A central thesis of this lexical approach is that the entire domain of job performance can be captured by culling all job performance measures used in the extant literature. This parallels the lexical hypothesis used in the personality literature which, as first enunciated by Goldberg, holds that a comprehensive description of the personality of an individual can be obtained by examining the adjectives used in the lexicon (e.g., all English language words that could be obtained/culled from a dictionary).

Viswesvaran (1993) listed job performance measures (486 of them) used in published articles over the years. Two raters working independently then derived 10 dimensions by grouping conceptually similar measures. The 10 dimensions were: overall job performance, job performance or productivity, effort, job knowledge, interpersonal competence, administrative competence, quality, communication competence, leadership, and compliance with rules. Overall job performance captured overall effectiveness, overall work reputation, or was the sum of all individual dimensions rated. Job performance or productivity included ratings of quantity or ratings of volume of work produced. Ratings of effort were statements about the amount of work an individual expends in striving to do a good job. Interpersonal competence was assessments of how well an individual gets along with others whereas administrative competence was a ratings measure of the proficiency exhibited by the individual in handling the coordination of the different roles in an organization. Quality was an assessment of how well the job was done and job knowledge was a measure of the expertise demonstrated by the individual. Communication competence reflected how well an individual communicated regardless of the content. Leadership was a measure of the ability to successfully bring out extra performance from others, and compliance with or acceptance of authority assessed the perspective the individual has about rules and regulations. Illustrative examples as well as more elaborate explanations of these dimensions are provided in Viswesvaran et al. (1996).

Although the lexical approach is promising, it should be noted that there are two potential concerns

here. First, it can be argued that just as the technical nuances of personality may not be reflected in the lexicon, some technical but important aspects of job performance have never been used in the literature – thus, not covered in the 10 dimensions identified. Second, it should be noted that generating 10 dimensions from a list of all job performance measures used in the extant literature involved the judgmental task of grouping conceptually similar measures.

Of these two concerns, the first is mitigated to the extent that the job performance measures found in the extant literature were identified by industrial-organizational psychologists and other professionals (in consultation with managers in organizations). As such the list of measures can be construed as a comprehensive specification of the entire domain of the construct of job performance. The second concern, the judgmental basis on which the job performance measures were grouped into 10 conceptual dimensions, is mitigated to the extent that intercoder agreement is high (the intercoder agreement in grouping the conceptually similar measures into the 10 dimensions was reported in the 90%s, Viswesvaran, 1993).

A comprehensive specification of the job performance construct involves many measures, the intercorrelation among which is needed to conduct the factor analyses. Estimating the correlations among all variables with adequate sample sizes may not be feasible in a single study. Fortunately, meta-analysis can be used to cumulate the correlations across pairs of variables, and the meta-analytically constructed correlation matrix can be used in the factor analyses (cf. Viswesvaran & Ones, 1995). Conway (1999) developed a taxonomy of managerial behavior by meta-analytically cumulating data across 14 studies, and found a three-level hierarchy of managerial performance. Viswesvaran (1993) cumulated results from over 300 studies that reported correlations across the 10 dimensions. Both interrater and intrarater correlations, as well as nonratings-based measures were analyzed. The 10 dimensions showed a positive manifold of correlations, suggesting the presence of a general factor across the different dimensions (Campbell, Gasser & Oswald, 1996).

CRITERIA FOR ASSESSING THE QUALITY OF INDIVIDUAL JOB PERFORMANCE ASSESSMENTS

For over a century, researchers have grappled with the issues involved in assessment of individual job performance (cf. Austin & Villanova, 1992, for a summary). It is no wonder that several researchers have advanced criteria for evaluating these assessments. Freyd (1926) argued that measures of individual job performance assessments should be validated. While Freyd argued for the importance of establishing the construct validity of criteria, Farmer (1933) stressed the need for assessing the reliability of measures. Burtt (1926) provided a list of variables (e.g., opportunity bias) that could affect organizational records or objective performance. Brogden and Taylor (1950) discussed the different types of criterion bias, specifically differentiating between bias that is correlated with predictor variables and biases that are unrelated to predictors.

Bellows (1941) identified six criteria that he grouped into statistical, acceptability, and practical effects categories. Bechtoldt (1947) introduced three criteria: (1) reliability and discriminability, (2) pertinence and comprehensiveness, and (3) comparability. Reliability is the consistency of measurement (Nunnally, 1978) and a good measure of assessment of individual job performance should discriminate across individuals. Pertinence refers to job-relatedness, and comprehensiveness requires that all important aspects of job performance are included in the assessment. Comparability focuses on the equivalence across the different dimensions assessed (e.g., time, place).

Thorndike (1949) proposed four criteria: (1) relevance, (2) reliability, (3) freedom from discrimination, and (4) practicality. Relevance is the construct validity of the measures, and can be construed as the correlation between the true scores and the construct (i.e., job performance). Given that this correlation can never be empirically estimated, relevance or construct validity is assessed by means of a nomological net of correlations with several related measures (see section on construct validity in Chapter 2 by Aguinis et al., in this volume). Relevance is the lack of criterion contamination (the measure includes what it should not include) and criterion deficiency (measure lacks what it should include). Note that Thorndike's use of the term 'discrimination' differs from use of the term by Bechtoldt (1947). For Thorndike discrimination is unfair distinctions made based on (demographic) group memberships. All measures designed to assess individual job performance should discriminate – the question is whether the discrimination is relevant to job performance or is unrelated to it.

Ronan and Prien (1966) argued that reliability of assessments is the most important factor in evaluating the quality of individual job performance assessments. Guion (1976), on the other hand, stressed the importance of assessing the construct validity of the performance assessments. Smith (1976) identified relevance (construct validity), reliability, and practicality as criteria for evaluating job performance assessments. Blum and Naylor (1968) summarize the conclusions of many researchers on criteria. Across the different classifications, the common criteria can be stated as (1) discriminability across individuals, (2) practicality, (3) acceptability, (4) reliability, (5) comprehensiveness (lack of

criterion deficiency), and (6) construct validity (or relevance or job relatedness or pertinence or freedom from bias such as contamination).

Of these six criteria, voluminous research has focused on issues of reliability and construct validity. Methods to assess job relatedness (pertinent) has been covered in other chapters (see Chapter 4 by Sanchez & Levine, this volume). Criteria such as discriminability and practicality pertain to administration issues and may depend on the context. For example, how well an individual counts can be a good measure of job performance for entry-level clerks in a grocery store but not for high-level accountants. Finally, there has been some limited research on user acceptability as a criterion. In light of this, I devote the rest of this section to these two issues – reliability and construct validity of individual job performance assessments – after briefly summarizing the research on user acceptability.

User Acceptability

Some recent research in the past 20 years has focused on user acceptability of peer ratings of individual job performance. Researchers (e.g., King, Hunter & Schmidt, 1980) have noted that raters were unwilling to accept nontransparent rating instruments such as the mixed standard scales and forced choice measures. Bobko and Colella (1994) summarize the research on how users make meaning and set acceptable performance standards. Dickinson (1993) reviews several factors that could affect user reactions. Folger, Konovsky and Cropanzano (1992) present a due process model based on notions of organizational justice to explain user reactions. User reactions were more favorable when adequate notice was given by the organization about the performance assessment process, a fair hearing was provided, and standards were consistently applied across individuals. Peer ratings were more accepted when peers were considered knowledgeable and have had opportunity to observe the performance.

Earlier research by Borman (1974) had suggested that involvement in the development of rating scales produced more favorable user reactions. This is consistent with the idea that the ability to provide input into a decision process enhances perceptions of procedural justice. Notions of informational and interactional justice (see Chapter 8 in Volume 2 by Gilliland & Chan) also affect user reactions. Taylor, Tracey, Renard, Harrison and Carroll (1995) found that when rater–ratee pairs were randomly formed with some raters trained in due process components, ratees assigned to the trained raters expressed more favorable reactions even though their performance evaluations were more negative compared to the ratees assigned to untrained raters. Several researchers (e.g., Villanova, 1992) have advanced a stakeholder model that explicitly takes into account the values which underlie performance assessments.

Reliability of Individual Job Performance Assessments

Reliability is defined as the consistency of measurement (Nunnally, 1978; Schmidt & Hunter, 1996). Mathematically it can be defined as the ratio of true to observed variance, and depending on what part of observed variance is construed as true variance and what is construed as error variance we have different reliability coefficients (Pedhazur & Schmelkin, 1991; Schmidt & Hunter, 1996). The three major types of reliability assessments that pertain to individual job performance are (1) internal consistency, (2) stability estimates, and (3) interrater reliability estimates. These reliability estimates can be computed for either overall job performance assessments or for each dimension assessed. Some of these estimates (e.g., interrater) are applicable to only some methods of assessments (e.g., subjective evaluations such as ratings) whereas other types of reliability estimates (e.g., stability) are applicable to all methods of assessment (subjective evaluations such as ratings and organizational records).

Consider a researcher interested in assessing the dimension of interpersonal competence in individual job performance. The researcher could develop a list of questions that relate to interpersonal competence and require knowledgeable raters to evaluate individuals in each of the questions. Either an unweighted or weighted sum of the responses to all questions is taken as a measure of interpersonal competence. Now in considering the observed variance across individuals, each question has a specific or unique variance as well as a shared variance with other items. To estimate what proportion of the observed variance is common or shared across items, we employ measures of internal consistency. The most commonly used measure of internal consistency is Cronbach's alpha (Cronbach, 1951).

Internal consistency estimates are also appropriate when organizational records are used to assess individual job performance. If several operational measures of absenteeism are obtained and absenteeism is defined as the common or shared variance across these different operationalizations, then an estimate of internal consistency of organizational records can be computed.

Stability estimates can be obtained as the correlation between measures obtained at times 1 and 2. Here true performance is construed as what is common to both time periods. The greater the time interval, the more likely that true performance will change. Coefficients of stability can be assessed for both organizational records as well as for subjective evaluations such as ratings. With ratings, the same

rater has to evaluate the individual at both times of assessment; if different raters are used, stability estimates of ratings confound rater differences with temporal instability.

To estimate the extent to which two raters will agree in their ratings, the interrater reliability is assessed as the correlation between the ratings provided by two raters of the same group of individuals. In reality, different sets of two raters are used to estimate different individuals; under such circumstances the interrater correlation also takes into account rater leniency. Interrater reliability is less applicable with measures based on organizational records, unless the interest is on estimating how accurately the performance has been recorded (better designated as interobserver or intercoder or interrecorder reliability). Interrater reliability can be assessed for overall job performance assessments as well as for specific dimensions of individual job performance.

Interrater reliability can be assessed for different types of raters: supervisors, peers, subordinates, clients/customers. One question that could be raised is whether there are two 'parallel' supervisors. That is, to estimate interrater agreement among supervisors, we need ratings of a group of individuals from at least two supervisors. In many organizations we have only one 'true' supervisor and a second individual (perhaps the supervisor to the supervisor) is included to assess interrater reliability. It could be argued that these two sets of ratings are not parallel. Although conceptually sound, the evidence we review below for interrater reliability of job performance assessments shows that this is not the case. The interrater reliability for peer ratings is lower than that for supervisor ratings (and presumably there are parallel peers).

The different types of reliability estimates for job performance assessments were explained in terms of correlations. However, analysis of variance models can also be used (Hoyt & Kerns, 1999). In fact, generalizability theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972) has been used as a framework to assess the variance due to different sources. Depending on how error variance is conceptualized, different generalizability coefficients can then be proposed. Some researchers (e.g., Murphy & DeShon, 2000) have mistakenly argued that generalizability theory alone estimates these different reliability estimates. In reality, correlational methods and analysis of variance models based on classical measurement theory can be (and were) used to estimate the different reliability estimates (generalizability coefficients). There is not much difference across the different frameworks when properly estimated and interpreted (Schmidt, Viswesvaran & Ones, 2000).

Several studies that had evaluated individual job performance report internal consistency estimates. Consistent with the predominance of cross-sectional studies in the literature compared to longitudinal studies, fewer studies have estimated stability coefficients. Further, more reliability estimates have been reported for subjective evaluations such as ratings than for measures of organizational records. Rather than reviewing each study (which is impossible even in a book-length format), I will summarize the results of major studies and meta-analyses conducted on this topic.

Rothe (1978) conducted a series of studies to assess the stability of productivity measures for different samples of chocolate dippers, welders, and other types of workers. Hackett and Guion (1985) report the reliability of absenteeism measures. Accident measures at two different time periods have been correlated.

Viswesvaran et al. (1996) conducted a comprehensive meta-analysis cumulating results across studies reporting reliability estimates for peer and supervisor ratings. Coefficient alphas, stability estimates, and interrater reliability estimates were averaged separately. The reliability was reported both for assessments of overall job performance as well as for nine dimensions of performance. For supervisory ratings of overall job performance, coefficient alpha was .86, the coefficient of stability was .81, and interrater reliability was .52. It appears that the largest source of error variance was due to rater-specific variance. This finding compares with the generalizability estimates obtained by Greguras and Robie (1998) as well as meta-analysis of the generalizability studies by Hoyt and Kerns (1999).

The reliability estimates for supervisory ratings of different dimensions of job performance are also summarized in Viswesvaran et al. (1996). The sample size weighted mean estimates (along with total number of estimates averaged and total sample size across averaged estimates) are provided below. Interrater reliability estimates were .57 (k = 19, N = 2,015), .53 (k = 20, N = 2,171), .55 (k = 24, N = 2,714), .47 (k = 31, N = 3,006), and .53 (k = 20, N = 14,072), for ratings of productivity, leadership, effort, interpersonal competence, and job knowledge, respectively. Coefficient alphas for ratings of productivity, leadership, effort, interpersonal competence were .82, .77, .79, and .77, respectively.

Viswesvaran et al. (1996) also report the sample size weighted average reliability for peer ratings. For ratings of overall job performance, interrater reliability was .42 and coefficient alpha was .85. Reliabilities for peer ratings of leadership, job knowledge, effort, interpersonal competence, administrative competence, and communication competence are also reported (see Viswesvaran et al., 1996). Average coefficient alphas for peer ratings of leadership, effort, and interpersonal competence were .61, .77, and .61, respectively.

Viswesvaran et al. (1996) focused on peer and supervisor ratings, whilst recent studies have

explored the reliability of subordinate ratings, for example Mount (1984) as well as Mount, Judge, Scullen, Stysma and Hezzlett (1998). Interrater reliability of subordinate ratings have been found to vary between .31 to .36 for the various dimensions of performance. Scarce data exist for assessing the reliability of customer ratings of performance, and research in the new millennium should remedy this deficiency in the literature.

Further research should also explore the effects of contextual variables in reliability assessments. Churchill and Peter (1984) as well as Petersen (1994) investigated the moderating effects of 13 variables on the reliability estimates of different variables (including job performance). No strong moderator effects were found. Rothstein (1990) reported that the interrater reliability of supervisor ratings of job performance is moderated by the length of exposure the rater has to the ratees. Similar effects such as opportunity to observe should be explored for their effects on reliability estimates. However, these moderating variables can also be construed as variables affecting the construct validity of ratings. It is erroneous to argue that since several variables could potentially affect ratings, interrater reliability estimates do not assess reliability. Reliability is not validity and validity is not reliability (Schmidt et al., 2000). I now turn to a discussion of the construct validity of individual job performance assessments.

Construct Validity of Individual Job Performance Assessments

The construct validity of a measure can be conceptualized as the correlation between the true scores from the measures and the underlying construct (i.e., individual job performance). This correlation can never be empirically estimated, and several lines of evidence are analyzed to assess construct validity. A major component of construct validity is to assess the convergent validity between different methods of assessing the same construct. Heneman (1986) meta-analytically cumulated the correlation between subjective evaluations of job performance provided by supervisors with organizational records-based measures of individual job performance. Heneman (1986) cumulated results across 23 studies (involving a total sample of 3,718) and found a corrected mean correlation of .27 between supervisory ratings and organizational records. Heneman used a reliability estimate of .60 for supervisory ratings and a test–retest stability estimate of .63 for output measures. Using a value of .52 for the reliability of supervisory ratings results in a correlation of .29.

Heneman's (1986) analyses were updated by Bommer, Johnson, Rich, Podsakoff and Mackenzie (1995), who also introduced refinements to the estimation of the convergent validity. Bommer et al. (1995) computed composite correlations across conceptual replications and cumulated the composite correlations. Composite correlations are more construct-valid than average correlations (see Viswesvaran, Schmidt & Ones, 1994, for a mathematical proof). Bommer et al., estimated the convergence validity between supervisory ratings and organizational records as .39, a value that agrees with the correlations estimated by Viswesvaran (1993). Both Heneman (1986) and Bommer et al. (1995) concluded that rating format and rating scale do not moderate the convergent validity.

McEvoy and Cascio (1987) estimated the correlation between turnover and supervisory ratings of job performance as –.28. This estimate of –.28 was based on a cumulation of results across 24 studies involving 7,717 individuals. McEvoy and Cascio (1987) had used a reliability estimate of .60 for supervisory ratings; using an estimate of .52, results in a correlation of –.30. Bycio (1992) meta-analyzed the results across studies reporting a correlation between absenteeism and job performance. Across 49 samples involving 15,764 datapoints, the correlation was –.29. This estimate of –.29 averaged results across studies that used either time lost or frequency measures of absenteeism. When the cumulation was restricted to time lost measures of absenteeism, the correlation was –.26 (28 samples, 7,704 individuals); when restricted to frequency measures of absenteeism, the estimate was –.32 (21 samples, 8,060 individuals).

In addition to investigating the convergence between supervisory ratings of job performance and organizational records of (1) productivity, and (2) personnel data such as turnover and absenteeism, researchers have explored the overlap between organizational records of productivity and personnel data (e.g., absenteeism, promotions etc.). Bycio (1992) reports a correlation of .24 between organizational records of performance indices and absenteeism (23 samples, 5,204 individuals). The correlation was –.28 (11 samples, 1,649 individuals) when time lost measures of absenteeism were considered; with frequency-based measures of absenteeism (12 samples, 3,555 individuals) the meta-analyzed correlation was –.22.

The meta-analytic results summarized so far focused on supervisory ratings and on ratings of overall job performance. Viswesvaran (1993) reports correlations between organizational records of productivity and 10 dimensions of rated job performance. The convergent validity of ratings and records-based measures were analyzed for peers and supervisors. In general, the convergent validity was higher for supervisory ratings than they were for peer ratings. Organizational records seem to reflect the supervisory perspective more than the peer perspective.

The convergence among the different sources of ratings have been explored, and two reviews of this literature have been reported. Mabe and West (1982) presented the first review of this literature which

was subsequently updated by Harris and Schaubroeck (1988). Harris and Schaubroeck (1988) found a correlation of .62 between peer and supervisory ratings of overall job performance (23 samples, 2,643 individuals). The correlation between self and supervisor or peer ratings were much lower. Whilst Harris and Schaubroeck focused on overall ratings of job performance, Viswesvaran, Schmidt and Ones (2000) meta-analyzed the peer–supervisor correlations for overall as well as eight dimensions of job performance. Viswesvaran et al. (2000) reported a mean observed peer–supervisor correlation of .40, .48, .38, .34, .35, .36, .41, and .49, for ratings of productivity, effort, interpersonal competence, administrative competence, quality, job knowledge, leadership, and compliance with authority, respectively. Research suggests (e.g., Harris, Smith & Champagne, 1995) that ratings obtained for administrative and research purposes are comparable.

Most of the extant literature reported correlations between self ratings, peer ratings, supervisor ratings, and organizational records. Recent research has started exploring the convergence between other sources of ratings (e.g., subordinates, customers). Mount et al. (1998) report correlations between subordinate ratings and peer or supervisor ratings for overall as well as three dimensions of performance. More research is needed in the future to make robust conclusions of convergent validity across these sources.

In addition to investigating the convergent validity across sources with correlations, researchers have used the multitrait–multimethod matrix (Campbell & Fiske, 1959) of correlations between different methods and performance dimensions to tease out the trait and method variance. Cote and Buckley (1987) as well as Schmitt and Stults (1986) provide elaboration of this approach as well as a summary of the application to performance assessment. Conway (1996) used an MTMM matrix and confirmatory factor analyses to support the construct validity of task and contextual performance measures. Mount et al. (1998), however, caution that previous applications of this approach had neglected within-source variability, and once this source is taken into account substantive conclusions vary.

Convergence across sources is one aspect of construct validity. Assessment of construct validity also involves assessing the potential and presence of several sources of variance that is unrelated to the construct under investigation. From as early as the 1920s researchers have been developing lists of factors that could affect the construct validity of job performance assessments. Burtt (1926) drew attention to the potential for criterion contamination and deficiency in organizational records, whilst Thorndike (1920) introduced the concept of halo error in ratings.

The last half of the twentieth century has seen an explosion of research on judgmental errors that could affect ratings. Lance, LaPointe and Stewart (1994) identified three definitions of halo error. Halo could be conceptualized as (1) a general evaluation that affects all dimensional ratings, (2) a salient dimension that affects ratings on other dimensions, and (3) insufficient discrimination among dimensions (Solomonson & Lance, 1997). Cooper (1981) discusses the different measures of halo as well as strategies designed to mitigate the effects of halo. Distributional problems such as leniency, central tendency, and stringency have been assessed. Judgmental errors such as the fundamental attribution error, representativeness and availability heuristics, and contrast effects in assessments have been studied. Wherry and Bartlett (1982) present a model incorporating many of the potential influences on ratings. Recent methodological advances such as combining meta-analysis and structural equations modeling, meta-analysis and generalizability theory (Hoyt, 2000; Hoyt & Kerns, 1999), have enabled researchers to assess the effects of these judgmental processes on the construct validity of job performance assessments.

Finally, investigations of the construct validity of ratings have been explored by estimating the effects of demographic variables on assessments. Kraiger and Ford (1985) reported differences between racial groups of almost one half of a standard deviation unit. However, the Kraiger and Ford (1985) meta-analyses included laboratory-based experimental studies as well as field studies. More importantly, ratee ability was not controlled. Pulakos, White, Oppler and Borman (1989) found in a large sample study of job performance assessment in a military setting, that once ratee ability is controlled, the biasing effects of race are small. Similar findings were found with civilian samples of over 36,000 individuals across 174 jobs (Sackett & Dubois, 1991). The effects of age and gender of the ratees and raters have also been investigated (see Cascio, 1991, for a summary). The biasing effects of demographic variables has not been found to be substantial. The dynamic nature of criteria has also been investigated, and empirical evidence suggests that although mean levels of individual job performance changed over time, rank ordering of individuals did not (Barrett, Cladwell & Alexander, 1989). Although potential exists for distortion, most well-constructed and administered performance assessments systems result in construct-valid data on individual job performance.

CAUSAL MODELS FOR JOB PERFORMANCE DIMENSIONS

In the last section of this chapter, I review models of work behavior that postulate how different individual

differences variables are linked to different aspects of performance. The search for explanation and understanding suggests a step beyond mere prediction (Schmidt & Kaplan, 1971). Hunter (1983) developed and tested a causal model where cognitive ability was a direct causal antecedent to both job knowledge and job performance. Job knowledge was an antecedent to job performance. Both job knowledge and job performance contributed to supervisory ratings. These findings suggest that cognitive ability contributes to overall job performance through its effects on learning job knowledge and mastery of required skills. Borman, Hanson, Oppler, Pulakos and White (1993) extended the model to explain supervisory performance.

McCloy, Campbell and Cudeck (1994) argued that all individual differences variables affect performance in any dimension by their effects on either procedural knowledge or declarative knowledge or motivation. Barrick, Mount and Strauss (1993) tested and found support for a model where conscientiousness predicted overall performance by affecting goal setting. Ones and Viswesvaran (1996) argued that conscientiousness has multiple pathways by which it affects overall performance. First, conscientious individuals are likely to spend more time on the task and less time daydreaming. This investment of time will result in greater acquisition of job knowledge, which in turn will result in greater productivity and which in turn will result in positive ratings. Further, conscientious individuals are likely to engage in organizational citizenship behaviors which in turn might enhance productivity and ratings. Finally, conscientious individuals are expected to pay more attention to detail and profit more via vicarious learning (Bandura, 1977) which would result in higher job knowledge and productivity.

Borman and Motowidlo (1993) postulated that ability will predict task performance more strongly than individual differences in personality. On the other hand, individual differences in personality were hypothesized to predict contextual performance better than ability. Motowidlo et al. (1997) developed a more nuanced model where contextual performance was modeled as dependent on contextual habits, contextual skills, and contextual knowledge. Although habits and skills were predicated on personality, contextual knowledge was influenced both by personality and cognitive ability. Similarly, task performance is influenced by task habits, task skill and task knowledge. Whereas task skill and task knowledge are influenced solely by cognitive ability, task habits are affected by both cognitive ability and personality variables. Thus, this more nuanced model implies that both ability and personality have a role in explaining task and contextual performance. The bottom line appears to be that each performance dimension is complexly determined so that it is impossible to specify different individual differences variables as sole cause or antecedent of a particular dimension of job performance. This is also to be expected given the positive correlations across the various dimensions.

CONCLUSIONS

Job performance is a central construct in our field. Voluminous research has been undertaken to assess (1) the factor structure of the construct, (2) refine the methods of assessment, (3) assess user reactions, reliability, and construct validity of assessments of individual job performance, and (4) develop models of work behavior that delineate the antecedents of individual job performance. A century of research suggests that the factor structure of job performance can be summarized as a hierarchy with a general factor at the apex with group factors at the next level. The breadth and range of the group factors differ across authors.

Several methods of assessments have been proposed, evaluated, and used. Research on user reactions has invoked justice theory concepts. Interrater reliability, internal consistency estimates, and stability assessments have been examined for assessments of overall performance as well as for several dimensions of performance. Correlational, Anova and generalizability models have been used in reliability estimation. The construct validity of individual job performance assessment has been assessed with emphasis on judgmental errors such as halo, group differences, convergences between different methods of assessments. Finally, path models have been specified to link antecedents to the different job performance dimensions.

Impressive as the existing literature is on assessments of individual job performance, several trends in the workplace call for additional research. The changing nature of work (Howard, 1996) brings with it the changes in assessment of performance (Ilgen & Pulakos, 1999); and the use of electronic monitoring and other technological advances may change the nature of what we measure (Hedge & Borman, 1995). Assessments of performance of expatriates will also gain in importance (see Chapter 20 by Sinangil & Ones in this volume). In short, a lively phase is ahead for researchers and practitioners.

REFERENCES

Alonso, A. (2000). The relationship between cognitive ability, the Big Five, Task and Contextual Performance: A meta-analysis. Unpublished Masters Thesis, Florida International University, Miami, FL.

Arvey, R.D., & Murphy, K.R. (1998). Performance evaluation in work settings. *Annual Review of Psychology*, *49*, 141–168.

Austin, J.T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology*, *77*, 836–874.

Baier, D.E., & Dugan, R.D. (1957). Factors in sales success. *Journal of Applied Psychology*, *41*, 37–40.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.

Barber, A.E. (1998). *Recruiting employees: Individual and organizational perspectives*. Thousand Oaks, CA: Sage.

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

Barrick, M.R., Mount, M.K., & Strauss, J. (1993). Conscientiousness and performance of sales representatives: Test of the mediating effects of goal setting. *Journal of Applied Psychology*, *78*, 715–722.

Bechtoldt, H.P. (1947). Factorial investigation of the perceptual-speed factor. *American Psychologist*, *2*, 304–305.

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

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

Bernardin, J.H., Alvares, K.M., Cranny, C.J. (1976). A recomparison of behavioral expectation scales to summated scales. *Journal of Applied Psychology*, *61*(5), 564–570.

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

Bernardin, H.J., Pence, E.C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology*, *65*(1), 60–66.

Bernardin, H.J., & Russell, J.E.A. (1998). *Human resource management: An experiential approach* (2nd ed.). Boston, MA: McGraw-Hill.

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

Blum, M.L., & Naylor, J.C. (1968). *Industrial Psychology: Its theoretical and social foundations*. New York: Harper & Row.

Bobko, P., & Colella, A. (1994). Employee reactions to performance standards: A review and research propositions. *Personnel Psychology*, *47*, 1–29.

Bommer, W.H., Johnson, J.L., Rich, G.A., Podsakoff, P.M., & MacKenzie, S.B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology*, *48*, 587–605.

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

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, CA: Consulting Psychologists Press.

Borman, W.C., Hanson, M.A., Oppler, S.H., Pulakos, E.D., & White, L.A. (1993). Role of early supervisory experience in supervisor performance. *Journal of Applied Psychology*, *78*, 443–449.

Borman, W.C., & Motowidlo, S.J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W.C. Borman (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco, CA: Jossey-Bass.

Borman, W.C., Motowidlo, S.J., Rose, S.R., & Hansen, L.M. (1985). *Development of a model of soldier effectiveness*. Minneapolis, MN: Personnel Decisions Research Institute.

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.

Brief, A.P., & Motowidlo, S.J. (1986). Prosocial organizational behavior. *Academy of Management Review*, *11*, 710–725.

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

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

Bycio, P. (1992). Job performance and absenteeism: A review and meta-analysis. *Human Relations*, *45*, 193–220.

Campbell, J.P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, 2nd ed., pp. 687–731). Palo Alto, CA: Consulting Psychologists Press.

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

Campbell, J.P., Gasser, M.B., & Oswald, F.L. (1996). The substantive nature of job performance variability. In K.R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 258–299). San Francisco: Jossey-Bass.

Campbell, J.P., McCloy, R.A., Oppler, S.H., & Sager, C.E. (1993). A theory of performance. In N. Schmitt & W.C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco, CA: Jossey-Bass.

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

Cascio, W.F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Church, A.H., & Bracken, D.W. (1997). Advancing the state of the art of 360 degree feedback. *Group and Organization Management*, *22*, 149–161.

Churchill, G.A., Jr., & Peter, J.P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research, 21*, 360–375.

Clark, J.P., & Hollinger, R.C. (1983). *Theft by employees in work organizations: Executive summary*. Washington, DC: National Institute of Justice.

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.

Conway, J.M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management*, *22*, 139–162.

Conway, J.M. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology*, *84*, 3–13.

Cooper, W.H. (1981). Ubiquitous halo. *Psychological Bulletin*, *90*(2), 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.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

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

DeVries, D.L., Morrison, A.M., Shullman, S.L., & Gerlach, M.L. (1986). *Performance appraisal on the line*. Greensboro, NC: Center for Creative Leadership.

Dickinson, T.L. (1993). Attitudes about performance appraisal. In H. Schuler, J.L. Farr & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 141–162). Hillsdale, NJ: Erlbaum.

DuBois, C.L., Sackett, P.R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and white-black differences. *Journal of Applied Psychology*, *78*, 205–211.

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

Folger, R., Konovsky, M.A., & Cropanzano, R. (1992). A due process metaphor for performance appraisal. In B.M. Staw & L.L. Cummings (Eds), *Research in Organizational Behavior* (Vol. 14, pp. 129–177). Greenwich, CT: JAI Press.

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

George, J.M., & Brief, A.P. (1992). Feeling good–doing good: A conceptual analysis of the mood at work-organizational spontaneity relationship. *Psychological Bulletin*, *112*, 310–329.

Goldberg, L.R. (1995). What the hell took so long? Donald Fiske and the big-five factor structure. In P.E. Shrout & S.T. Fiske (Eds.), *Advances in personality research, methods, and theory: A festschrift honoring Donald W. Fiske*. New York, NY: Erlbaum.

Greguras, G.J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology*, *83*, 960–968.

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. (1998). *Assessment, measurement, and prediction for personnel selection*. Mahwah, NJ: Lawrence Erlbaum.

Gunderson, E.K.E., & Ryman, D.H. (1971). Convergent and discriminant validities of performance evaluations in extremely isolated groups. *Personnel Psychology*, *24*, 715–724.

Hackett, R.D., & Guion, R.M. (1985). A re-evaluation of the absenteeism-job satisfaction relationship. *Organizational Behavior and Human Decision Processes*, *35*, 340–381.

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

Harris, M.M., Smith, D.E., & Champagne, D. (1995). A field study of performance appraisal purpose: Research- versus administrative-based ratings. *Personnel Psychology*, *48*, 151–160.

Hedge, J.W., & Borman, W.C. (1995). Changing conceptions and practices in performance appraisal. In A. Howard (Ed.), *The changing nature of work* (pp. 451–481). San Francisco, Jossey-Bass.

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

Howard, A. (Ed.) (1996). *The changing nature of work*. San Francisco, Jossey-Bass.

Hoyt, W.T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*, 64–86.

Hoyt, W.T., & Kerns, M.D. (1999). Magnitude and moderators of bias in observer ratings: A meta analysis. *Psychological Methods*, *4*, 403–424.

Hunt, S.T. (1996). Generic work behavior: An investigation into the dimensions of entry-level, hourly job performance. *Personnel Psychology*, *49*, 51–83.

Hunter, J.E. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization to General Aptitude Test Battery* (USES Test Research Report no. 45). Washington, DC: United States Department of Labor.

Ilgen, D.R., & Pulakos, E.D. (1999). *The changing nature of performance: Implications for staffing, motivation, and development*. San Francisco: Jossey-Bass.

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

Katz, D. (1964). The motivational basis of organizational behavior. *Behavioral Science*, *9*, 131–146.

Kidwell, R.E., & Bennett, N. (1993). Employee propensity to withhold effort: A conceptual model to intersect three avenues of research. *Academy of Management Review*, *18*, 429–456.

King, L.M., Hunter, J.E., & Schmidt, F.L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology*, *65*, 507–516.

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

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

Lance, C.E., LaPointe, J.A., & Stewart, A.M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology*, *79*, 332–340.

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

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

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.

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.

McCloy, R.A., Campbell, J.P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology*, *79*, 493–505.

McEvoy, G.M., & Cascio, W.F. (1987). Do good or poor performers leave? A meta-analysis of the relationship between performance and turnover. *Academy of Management Journal*, *30*, 744–762.

Motowidlo, S.J., Borman, W.C., & Schmit, M.J. (1997). A theory of individual differences in task and contextual performance. *Human Performance*, *10*, 71–83.

Mount, M.K. (1984). Psychometric properties of subordinate ratings of managerial performance. *Personnel Psychology*, *37*, 687–702.

Mount, M.K., Judge, T.A., Scullen, S.E., Stysma, M.R., & Hezlett, S.A. (1998). Trait, rater, and level effects in 360-degree performance ratings. *Personnel Psychology*, *51*, 557–576.

Murphy, K.R. (1989). Dimensions of job performance. In R. Dillon & J. Pelligrino (Eds.), *Testing: Applied and theoretical perspectives* (pp. 218–247). New York: Praeger.

Murphy, K.R., & Cleveland, J.N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Murphy, K.R., & DeShon, R. (2000). Inter-rater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873–900.

Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw Hill.

Ones, D.S., & Viswesvaran, C. (1996). A general theory of conscientiousness at work: Theoretical underpinnings and empirical findings. In J.M. Collins (Chair), *Personality predictors of job performance: Controversial issues*. Symposium conducted at the eleventh annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA, April.

Ones, D.S., Viswesvaran, C., & Schmidt, F.L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*, 679–703.

Organ, D.W. (1988). *Organizational citizenship behavior*. Lexington, MA: D.C. Heath.

Pedhazur, E.J., & Schmelkin, L.P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Peterson, R.A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, *21*, 381–391.

Prien, E.P., & Kult, M. (1968). Analysis of performance criteria and comparison of a priori and empirically-derived keys for a forced-choice scoring. *Personnel Psychology*, *21*, 505–513.

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.

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

Ronan, W.W. (1963). A factor analysis of eleven job performance measures. *Personnel Psychology*, *16*, 255–267.

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

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

Rothstein, H.R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, *75*, 322–327.

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

Sackett, P.R., & DuBois, C.L. (1991). Rater-ratee race effects on performance evaluations: Challenging meta-analytic conclusions. *Journal of Applied Psychology*, *76*, 873–877.

Sackett, P.R., & Wanek, J.E. (1996). New developments in the use of measures of honesty, integrity, conscientiousness, dependability, trustworthiness and reliability for personnel selection. *Personnel Psychology*, *49*, 787–830.

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

Schmidt, F.L. (1980). The measurement of job performance. Unpublished manuscript.

Schmidt, F.L., & Hunter, J.E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology*, *74*, 368–370.

Schmidt, F.L., & Hunter, J.E. (1992). Causal modeling of processes determining job performance. *Current Directions in Psychological Science*, *1*, 89–92.

Schmidt, F.L., & Hunter, J.E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199–223.

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

Schmidt, F.L., Viswesvaran, C., & Ones, D.S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*, 901–912.

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

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

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

Smith, C.A., Organ, D.W., & Near, J.P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology*, *68*, 655–663.

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.

Solomonson, A.L., & Lance, C.E. (1997). Examination of the relationship between true halo and halo error in performance ratings. *Journal of Applied Psychology*, *82*, 665–674.

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

Taylor, M.S., Tracey, K.B., Renard, M.K., Harrison, J.K., & Carroll, S.J. (1995). Due process in performance appraisal: A quasi-experiment in procedural justice. *Administrative Science Quarterly*, *40*, 495–523.

Thorndike, E.L. (1920). A constant error in psychological ratings. In J.P. Porter & W.F. Book (Eds.), *The Journal of Applied Psychology*, *4*, 25–29.

Thorndike, R.L. (1949). *Personnel selection: Test and measurement techniques*. New York: Wiley.

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

Van Dyne, L., Cummings, L.L., & Parks, J.M. (1995). Extra-role behaviors: Its pursuit of construct and definitional clarity (a bridge over muddied waters). In L.L. Cummings & B.M. Staw (Eds.), *Research in organizational behavior* (Vol. 17, pp. 215–285). Greenwich, CT: JAI Press.

Van Scotter, J.R., & Motowidlo, S.J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology*, *81*, 525–531.

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

Viswesvaran, C. (1993). Modeling job performance: Is there a general factor? Unpublished doctoral dissertation, University of Iowa, Iowa City, IA.

Viswesvaran, C., & Ones, D.S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology*, *48*, 865–885.

Viswesvaran, C., & Ones, D.S. (2000). Perspectives on models of job performance. *International Journal of Selection and Assessment*, *8*, 216–226.

Viswesvaran, C., Ones, D.S., & Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574.

Viswesvaran, C., Schmidt, F.L., & Ones, D.S. (1994). Examining the validity of supervisory ratings of job performance using linear composites. Paper presented in F.L. Schmidt (Chair), *The construct of job performance*. Symposium conducted at the ninth annual meeting of the Society of Industrial and Organizational Psychologists, Nashville, Tennessee, April.

Viswesvaran, C., Schmidt, F.L., & Ones, D.S. (2000). The moderating influence of job performance dimensions on convergence of supervisory and peer ratings of job performance: Unconfounding construct-level convergence and rating difficulty. Unpublished manuscript.

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

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

Welbourne, T.M., Johnson, D.E., & Erez, A. (1998). The role-based performance scale: Validity analysis of a theory-based measure. *Academy of Management Journal*, *41*, 540–555.

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

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

Whisler, T.L., & Harper, S.F. (Eds.) (1962). *Performance appraisal: Research and practice*. New York: Holt.