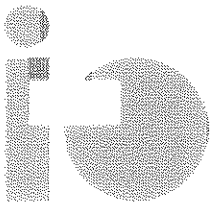


CHAPTER 5

Job Behavior, Performance, and Effectiveness

Walter C. Borman
University of South Florida

This chapter is about criterion development and criterion measurement. Major topics include important views and concepts related to criteria, recent developments in methods of measuring criterion performance, conceptual and methodological advantages and disadvantages of various kinds of criterion measures, and conclusions about criterion development and measurement. Because performance ratings are so often used as criteria in personnel research applications, considerable attention is focused on ratings. Areas discussed include design and evaluation of rating forms, analysis of ratings from different sources (e.g., peers or supervisors), identification of training treatments to reduce rating error and enhance the accuracy of performance ratings, examination of the performance judgment process to aid in understanding and improving ratings, and analysis of strategies for evaluating performance ratings, including assessment of psychometric properties, interrater reliability, convergent and discriminant validity, and accuracy. Also critically reviewed as criteria are turnover, absences, production rates, job level and salary, sales, disciplinary cases, performance tests or work samples, and job knowledge tests. Parallel to what has been accomplished with predictor domains such as cognitive abilities and personality, the development and evaluation of criterion models is offered as a promising direction for criterion research. Structural and latent variable modeling may be used to identify and confirm performance constructs in order to enhance the scientific understanding and usefulness of criteria.

MEASURES OF CRITERION performance are necessary for almost all competently conducted personnel research applications in organizations. If we are to assess empirically the impact of any personnel action on individual or group performance, criteria are essential. Criteria are needed to assess the effectiveness of personnel selection procedures, organizational training programs, job design efforts, and many other personnel-related actions and interventions. In personnel selection, for example, we may develop an experimental selection test, administer it to applicants for a job, and later assess the performance of those hired to evaluate the test's validity by correlating test scores with performance scores. This naturally requires some measure of performance that fairly and accurately depicts actual performance levels. In sum, criterion measures have great importance for practical applications in personnel research.

Criteria and criterion measures are, however, of considerable interest in their own right. Much has been made of learning as much as possible about what tests measure and how test scores should be interpreted, based especially on the scientific principles of construct validation (Cronbach & Meehl, 1955). A similar framework is appropriate for learning about criterion measures. To echo the work of others such as James (1973) and P. C. Smith (1976), researchers need to apply the same degree of effort and rigor to develop criteria and criterion measures that they used to develop predictor tests.

Views and Observations on Criteria

In this section, we review and discuss several important issues relevant to criteria and criterion measurement.

The Concept of Criterion Relevance and Other "Criteria for Criteria"

The most important standard for criteria is relevance. *Relevance* refers to the correspondence between criteria and the actual performance demands of the target job. A criterion measure should assess one or more of the job's important performance requirements; as a set, criterion measures should provide comprehensive coverage of all important performance requirements of the job.

The terms "contamination" and "deficiency" have been useful in assessing criterion relevance. *Contamination* is said to exist when a criterion measure taps variance irrelevant to the performance requirements. For example, a sales-per-month criterion for a computer software salesperson may be a function of this person's sales ability *and* the ease of software sales in his or her region. This measure is contaminated. Criterion *deficiency* becomes an issue when a set of criteria for a job fails to measure one or more of its important performance areas. For example, a work sample performance test for a typist position may very faithfully reflect technical proficiency on that job but completely fail to tap important interpersonal requirements. This measure is therefore deficient. A totally relevant set of criterion measures is thus neither contaminated nor deficient.

In practice, it may be possible to correct for criterion contamination where the source and degree of contamination can be identified, such as in adjusting for unequal opportunity to perform effectively, but nothing can be done about deficiency, short of obtaining or creating additional criterion measures. This issue will be discussed in more detail later in this chapter.

Early writings about criteria for criteria describe several standards in addition to relevance. A brief discussion of these historic

statements should be useful here, not only to understand standards for criteria themselves, but to help provide a sense of the history of criterion development and measurement.

Some time ago, Bellows (1941) listed accessibility and cost, acceptability to the sponsor, and predictability as three important criteria for criteria. The first two reflect an emphasis on the practical, use-what's-available approach to criterion measurement, an approach many (e.g., P. C. Smith, 1976) have argued against. The third criterion illustrates the logical fallacy of selecting criteria in personnel selection research according to the magnitude of predictor-criterion relationships, no matter what the particular nature of the criteria (e.g., their relevance) may be. This is a classic case of misunderstanding the criterion development process and ignoring the issue of relevance in criterion measurement.

Some time later, Toops (1944) advanced a relatively sophisticated treatise on criteria for criteria and related problems. He provided an in-depth analysis of criterion contamination, including such problems as the effects of teamwork on individual productivity, the issue of equal pay for both sexes based on equal levels of productivity, and problems with environmental constraints, such as relatively slow machine speed affecting criterion scores (production in this case).

Jenkins (1946), reviewing then-recent advances in applied testing practices, noted that from 1920 through 1940 considerable progress was made on the predictor development side, but almost no attention was given to criteria. He observed that during the twenties and thirties much was learned about the validation of tests, *given a criterion*, but that was exactly the problem: Criteria were considered "given of God or just to be found lying around" (p. 93). Jenkins then discussed advances in thinking about and working with criterion measures on the part of American psychologists during World War II, including quite a sophisticated discussion of such problems as criterion unreliability and deficiency, low correlations between training and job performance, and the sometimes dynamic nature of job performance requirements over time.

Fiske (1951) argued that an "ideal" approach to criterion development was to determine the contributions of each criterion behavior to the goals of the organization in assessing individual performance, urging that empirical research be used instead of "value judgments" in developing criteria. It is unclear how values about effectiveness can be divorced from operational definitions of performance.

Wherry (1957) referred to criteria for criteria by observing that psychometricians (like himself) should be depressed about the state of the art in criterion development compared to predictor test development. This was an early call for more rigorous criteria for criterion measures.

Weitz (1961), in his often-cited criteria for criteria paper, listed time, type, and level as "criterional dimensions" to be considered in developing and better understanding criterion measures. *Time* refers to when the criterion measure is taken, *type* has to do with the specific kind of criterion measure selected, and *level* pertains to the cutoff score for acceptable or unacceptable performance. Level is not actually important if criterion scores are treated as continuous variables, as they often are.

Wallace (1965) argued that the criterion for criteria of relevance to the total job might be less important than relevance to a particular research hypothesis that would lead to greater understanding of predictor-criterion relationships. He offered an example in the life insurance industry of considering as criteria total sales, which may be very relevant to the overall performance of an insurance salesperson, *or* a measure of number of calls made to potential clients (per unit time), which is presumably

a behavioral element contributing to sales success but perhaps not as relevant to overall sales performance. If we had a test that was hypothesized to predict this latter call-willingness/reluctance criterion, then Wallace's suggestion is to utilize this criterion because it is most pertinent to the predictor-criterion hypothesis.

The Ultimate Criterion Model

Thorndike (1949) proposed what is essentially a *hypothetical criterion construct*. For each situation in which criteria are required, we might conceive of an ultimate criterion, a single measure that would optimally summarize all relevant performance requirements in that situation. If such a criterion could be developed, it would presumably be a weighted linear composite of all important criterion elements. Thorndike offers an example: A person displaying maximum performance on the ultimate criterion for an insurance sales job might be one who sells the maximum amount of insurance it is possible to sell, allows none of these policies to lapse, and continues at this maximum performance level for many years.

The ultimate criterion concept can best be thought of as hypothetical and as a special case of composite criteria. More will be said about composite criteria in the next subsection.

Multiple and Composite Criteria

We might ask the following question about criteria: For the typical job, should we have a single criterion measure, or is it more appropriate to identify multiple criteria? Supporters can be found for both positions. Advocates of a composite criterion (e.g., Brogden & Taylor, 1950; Nagel, 1953) view the criterion as basically economic in nature, whereas those favoring multiple criteria (e.g., Dunnette, 1963; Guion, 1961; P. C. Smith, 1976; Wallace, 1965) believe criteria should represent behavioral or psychological constructs.

The central issue is to identify the purpose of criterion measurement (Schmidt & Kaplan, 1971). In making personnel selection decisions, for example, it is necessary at some point to combine multiple criteria to form a composite. Qualities such as overall success, worth as an employee, and contribution to the organization must be determined in order to select persons with the highest predicted overall performance. If the goal is increased understanding of predictor-criterion links, then multiple criteria are more appropriate. Continuing with the selection example, if individual criterion elements refer to very different kinds of performance (e.g., technical and interpersonal), different predictors (e.g., ability tests for the former and temperament construct measures for the latter) are likely to correlate with performance in each of these areas. Combining such criteria in a composite masks relationships between individual predictors and criteria, relationships that could increase understanding of predictors, criteria, and the relationship between them.

In addition to the question of research goals and strategies bearing on the use of composite or multiple criteria, there is the empirical question of how multidimensional the criteria *are* for jobs. If criterion measures are highly correlated in a job, then combining them to form a composite seems appropriate. Studies by Rush (1953), Seashore, Indik, and Georgopoulos (1960), Peres (1962), Ronan (1963), and others empirically demonstrate the multifactor nature of job performance. More recent work confirms the multidimensionality of job performance (e.g., J. P. Campbell, 1986).

This empirical question concerning the multidimensionality of criteria becomes complicated, however. Low correlations between criterion measures can be due to unreliability (Marks, 1967). In addition, method-specific variance may lead to low relationships between criteria tapped using different methods (e.g., work sample or knowledge test scores correlated with ratings). The method variance issue

is even more complex because method is easily confounded with the actual content of criterion elements. For example, job sample tests are said to tap maximum performance, the "can-do," technical proficiency aspects of job performance. Ratings, on the other hand, may reflect more the typical performance over time, "will-do" elements of performance. Accordingly, low relationships between such criterion indices might be a function of the different criterion constructs being focused on and the different methods being used. Thus, measurement issues including reliability of criterion measures and method variance associated with these measures when multiple methods are used cloud the empirical question regarding the multidimensionality of job performance. Interestingly, factor analyses of rating data, a *single method* in the context of this discussion, have also revealed multiple dimensions of job performance (cf. Pulakos, Borman, & Hough, 1988), which helps to confirm the multidimensional nature of performance.

Finally, it may be useful to consider the *expected correlation* between criterion elements in jobs. Cooper's (1981) thesis is relevant here. He argues that different dimensions of performance for individual jobs are likely to be correlated because of the way jobs are structured. A job usually requires incumbents to perform on a reasonably homogeneous set of tasks—that is, tasks that have similar knowledge, skill, and ability (KSA) requirements. If this were not the case, positions would be difficult to fill, with widely divergent KSAs necessary for successful performance. Because the KSA requirements are typically similar across tasks for a job, performance on the different dimensions of job performance will usually be correlated.

On balance, performance requirements for jobs are likely most faithfully represented by multiple criteria, with these criteria positively correlated to some extent. Later we will discuss the notion of developing models of job performance to reflect explicitly the nature and structure of multiple criterion constructs for a job or family of jobs. The model-building effort directly addresses this difficult problem of attempting to identify and then represent multiple criteria using fallible job performance measures.

Astin's Distinction Between Conceptual Criteria and Criterion Measures

Astin (1964) provided definitions and a useful discussion of different concepts pertinent to criteria. He defined the term *conceptual criterion* as a verbal statement of the important or socially relevant outcomes related to a particular problem. *Criterion measures*, in turn, refer to operational definitions of the conceptual criteria (e.g., performance rating scales or job sample tests). This distinction is important because it implies a logical sequence for identifying or developing criterion measures that are suitable for indexing criterion performance. First, conceptual criteria should be carefully identified to include all important dimensions of performance. This exercise will typically involve detailed articulation of the important performance requirements for the target job—articulations made with considerable help from the sponsor organization. *After* the conceptual criterion has been developed, attempts can be made to identify or develop operational measures related to each component of the conceptual criterion.

The ordering of these two steps is important. It prevents the common practice of using criterion measures simply because they are available or easily developed. It also ensures that, while criterion measures are being identified or developed, the researcher can assess the extent of conceptual criterion coverage provided by the criterion measures and thus estimate the deficiency of the measures. If a set of measures is very deficient in reflecting important conceptual criteria, then more work is

needed to increase this coverage. In fact, Astin argued that the validity of criterion measures can only be evaluated rationally—not empirically—by a logical analysis of these measures' relevance to the conceptual criteria.

The Campbell et al. Distinction Between Behavior, Performance, and Effectiveness

J. P. Campbell, Dunnette, Lawler, and Weick (1970) defined behavior as what managers actually do. Behaviors, with no evaluative component, might include tasks managers perform and activities that involve them. Performance criteria are developed by determining the value of behaviors to important organizational outcomes. Performance, then, reflects members' contributions to organizational goals—behaviors that lead to or detract from a position's contribution to organizational effectiveness. Performance criteria in this system are equivalent to Astin's *criterion measures;* they are operational definitions of important performance requirements that permit assessments of individual differences in performance levels.

Effectiveness has to do with outcomes. In the J. P. Campbell et al. model, global outcome measures such as promotion rate, salary level, and productivity indices are differentiated from performance measures because the effectiveness indices reflect not only the individual's contribution to effectiveness but also reflect factors beyond his or her control. J. P. Campbell et al. argued that the appropriate focus in their model for criterion development efforts is within the *performance* domain.

More recent conceptual and empirical work confirms the potential of external factors to influence the effectiveness of individuals in organizations. Peters and O'Connor (1980) hypothesized that constraints on performance, such as lack of proper tools, absence of required help from supervisors or co-workers, or insufficient job information, may lead to lower effectiveness levels and in turn reduce ability-effectiveness relationships. In a series of laboratory experiments, Peters, O'Connor, and colleagues demonstrated that constraints on performance can adversely affect outcome measures of effectiveness (Peters, Chassie, Lindholm, O'Connor, & Kline, 1982; Peters, O'Connor, & Rudolf, 1980). Field studies correlating severity of constraints with performance ratings show a reduced effect for constraints on performance (O'Connor, Peters, Rudolf, & Pooyan, 1982; Olson & Borman, 1989). In any case there is some support for the concern that outcome effectiveness measures are confounded in the sense that they reflect both the skill and effort of organization members *and* factors beyond their control.

Ghiselli's Concepts of Dynamic Criteria and Criterion Dimensionality

Ghiselli (1956) observed that the nature of criterion performance requirements may change over time as employees learn and develop on the job. He suggested that this *dynamic criterion* phenomenon could cause certain abilities or personal characteristics to be good predictors of performance at one point in an employee's tenure but not at another. For example, early in a salesperson's career, an aggressive search for clients may be important for success, while later on, maintaining warm and cordial relationships with customers might become more critical. Thus, it is conceivable that the rank order of salespersons' effectiveness could change with changes over time in the job's performance requirements.

Fleishman's laboratory experiments (Fleishman & Fruchter, 1960; Fleishman & Hempel, 1954) focused on a similar phenomenon. A major finding was that ability, psychomotor, and perceptual skill requirements for performing motor tasks change as learning progresses. The very nature of the task is transformed as learning changes the way persons approach performance on the task. With different

abilities and skills becoming necessary as time spent on the task increases, the rank order of subjects' performance levels changes, and the patterns of validity of these skill and ability predictors change over time. In particular, for novel perceptual-psychomotor tests, for example, general cognitive ability seems to better predict early task performance, while perceptual speed and psychomotor ability are better at predicting later performance (Ackerman, 1987).

However, much remains to be learned about the extent of this phenomenon in jobs and careers. Barrett, Caldwell, and Alexander (1985) reviewed 12 studies with data bearing on the existence (or not) of dynamic criteria and found little evidence for the phenomenon. They concluded that the dangers for personnel selection practice posed by dynamic criteria are not so serious as was feared, and that more concern should be focused on improving the reliability of criterion measures. The critical analysis of Barrett et al. is interesting, but the dynamic criterion concept remains of considerable theoretical and conceptual interest. Researchers should keep in mind the possibility that job requirements could change sufficiently over time to alter the patterns of validities for predictors of job performance.

Criterion dimensionality of the individual is another intriguing—and complicating—concept in the area of criterion measurement (Ghiselli, 1956). The notion is that two or more persons on the same job may be equally effective but may reach that level of performance very differently in behavioral terms. In a management job, for example, one manager may lead with charisma and flair, while another may incorporate a participative, caring style; both approaches can result in effective managerial performance. Thus, different dimensions of performance are relevant for assessing the effectiveness of these two managers, and different measures will likely be successful in predicting the performance of these two managers. In jobs where very different behavioral patterns are possible for success, this could be a significant criterion problem; however, research is needed to examine the extent of this phenomenon in actual organizational settings.

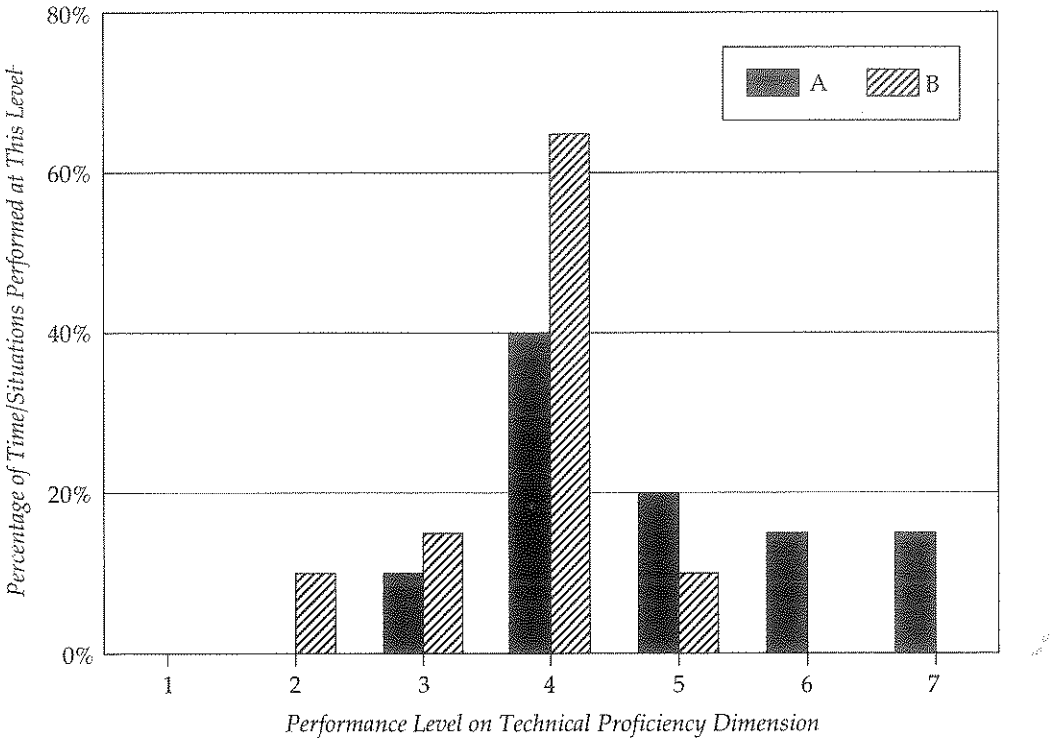
Evaluating Variability in Job Performance

Current instruments for performance rating provide estimates of *modal* performance. That is, rating forms require raters to estimate typical performance, essentially ignoring any variation in performance levels. From a measurement perspective, this translates into making judgments about a ratee's modal level of performance on each dimension. Yet it seems obvious that employees' performance on individual dimensions varies over time and across different situations on the job. In fact, such performance is probably more faithfully characterized by a distribution than it is by a single number. And if performance distributions for individual ratees could be accurately determined, considerable information appropriate for theoretical and practical purposes might be revealed. In addition to the mode, the variance and the skewness of the distribution, for example, could prove useful.

Consider the two performance distributions graphed in Figure 1. Although these employees have the same modal performance on this dimension, A performs at close to his or her minimum level most of the time, whereas B performs closer to his or her maximum level and is more consistent than A in performing on this dimension. In addition, it may be possible to draw inferences about the abilities and motivation of the two employees. A is capable of performing at the highest level but generally performs in the average range; B is perhaps more limited regarding capabilities in this aspect of the job but performs close to his or her own highest level most of the time.

FIGURE 1

Performance Distributions for Two Employees



An attempt has been made to develop a rating system that yields certain useful parameters of a ratee's performance distribution. Kane (1986) derived what he calls a *distributional measurement model*, in which variability in effort or motivation and external constraints beyond the control of the ratee are viewed as important and are addressed through his *performance distribution assessment method* (PDA).

With the PDA method, the rater is asked to record for each performance dimension the percentage of time that each level of performance cannot be attained because of factors beyond the ratee's control, then to note the percentage of time the ratee performed at or above each successive performance level. Kane (1986) provides formulas that allow computation of an average performance score, a consistency-in-performance score, and what he calls a *negative-range avoidance score*, an index of how successfully the ratee avoids poor performance. Thus, in Kane's system, variability in performance is explicitly addressed and indexed. Unfortunately, the usefulness of the system is not well known; for example, can raters make reliable judgments of the percentages? Nonetheless, the PDA system is a bold

initiative to obtain estimates of performance distribution parameters.

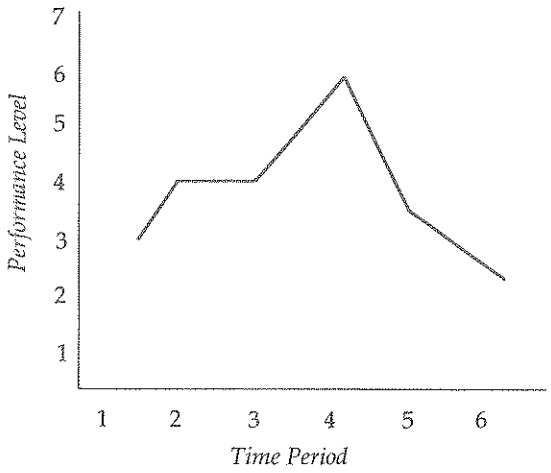
An alternative way to depict variability in job performance is to consider even more directly performance levels over time. Mapping performance levels of individuals or groups, as in time-series analyses (Glass, Willson, & Gottman, 1975; McCain & McCleary, 1979), allows a picture to emerge of modal performance and variability in performance. In addition, this method provides a display of performance slope—that is, whether performance levels are ascending, descending, or remaining steady, as is shown in Figure 2.

Komaki, Collins, and Penn (1982) used time-series analysis to evaluate safety performance in a food processing plant before and after training to decrease accident rates. Although these researchers were focusing on work group performance, the same strategy could be applied to individual performance. Of course, a major disadvantage of this method is that several measures of performance over time must be generated for individual employees.

In addition to the scientific merit of obtaining these more refined pictures of employee performance, the degree of variability or consistency of job performance is of considerable practical importance for some jobs. For example, a power plant operator should display very consistent high levels of performance on dimensions such as technical competence and decision making. A high modal level of performance with considerable variability is not acceptable. The point, then, is that performance distributions may provide comparatively rich descriptions of individuals' performance, giving us substantially greater understanding of that performance, along with its causes and consequences.

FIGURE 2

Time-Series Performance Data for an Employee



Recent Developments in "Enlarging the Criterion Space"

Equal employment opportunity considerations have led to increased concern that performance criteria be *job-related*, especially in personnel selection research (*Uniform Guidelines*, 1978). This is as it should be. However, a question might be raised about the boundaries of what is considered job-related. Some recent research and thinking has defined job performance much more broadly.

C. A. Smith, Organ, and Near (1983) have discussed what they call *organizational citizenship behavior* (see also Organ, 1988), which refers to behavior beyond task proficiency. In their conceptual and empirical work, they identified two factors: (a) altruism—day-to-day, prosocial behaviors toward others in the organization that help these others perform more effectively and (b) generalized compliance—"good-soldier," conscientiousness behaviors that demonstrate concern for doing things properly for the good of the organization. An intriguing finding is that job satisfaction–job performance relationships where performance is defined more broadly in the manner of the first

citizenship factor are substantially higher than satisfaction–performance correlations in general. A possible explanation is that high job satisfaction leads to more interest in, and tendency to carry out, altruistic prosocial initiatives at work.

Similarly, Borman, Motowidlo, Rose, and Hanser (1987), in identifying the performance requirements of enlisted soldiers in the U.S. Army, developed a model of soldier performance that included dimensions related to organizational commitment and socialization (e.g., adjustment to the Army and following orders and regulations). Their position was that such behavioral elements can be considered performance criteria as long as unit members' increased performance in these areas increases organizational effectiveness. It appears that citizenship, commitment, and socialization behaviors may contribute to an organization's effectiveness, especially in a team-oriented environment.

Related also to criterion areas beyond task proficiency, Brief and Motowidlo (1986) provide a comprehensive treatment of prosocial behavior in organizations—that is, behavior on the part of organization members that is intended to promote the welfare of the individual, group, or organization. Their discussion distinguishes between role-prescribed and extra-role activities in the prosocial domain. In some cases, this kind of behavior is explicitly required. For example, in the Borman et al. model, "supporting and providing guidance to other unit members" is one of the performance dimensions and is thus a role-prescribed behavior. In other cases, it is outside the formally *required* job role and is behavior that goes beyond the call of duty to help out an individual, a group, or the entire organization. Interestingly, Brief and Motowidlo point out that prosocial behavior may at times be dysfunctional to the organization, as when a unit member helps a co-worker with a personal problem and in the process fails to complete an important job-related task. Nonetheless, in most cases, organizational effectiveness is probably enhanced by prosocial behavior.

It seems clearer, however, that *task performance* is linked to organizational performance. More empirical research is needed to assess individual performance–organizational effectiveness relationships for these additional non–task-related elements of performance. If such links are found to be substantial, it would provide more justification for considering these less obvious elements as legitimate performance criteria for individual organization members.

Methods of Measuring Criterion Performance

Four major types of measures are used to assess criterion performance. The most often used are ratings—estimates of individuals' performance made by supervisors, peers, or others familiar with the incumbent's work behavior. Objective measures of performance or of other criterion behavior (e.g., absences or turnover) are also frequently used as criterion measures. Performance tests or job sample simulations are sometimes applied to measure task proficiency on jobs. Finally, written job knowledge tests are often used to evaluate success in training and, on occasion, the technical knowledge component of job performance. Descriptions of these measures and related research follow.

Performance Ratings

The emphasis in this section will be on ratings gathered for research only as criteria for personnel research applications. Although ratings can be generated for purposes of salary administration, promotion and termination decisions, or employee feedback and development, and although performance appraisal systems to address these administrative functions are

extremely important to individual and organizational effectiveness (cf. DeVries, Morrison, Shullman, & Gerlach, 1981), they are beyond the scope of this chapter.

Performance ratings are indeed the most often used criterion measure in industrial and organizational psychology. Landy and Farr (1980) refer to several surveys intended to assess how frequently ratings are used as criterion measures in research reports. The percentages reach 75 percent and higher, suggesting that considerable attention should be paid to this criterion measurement method.

Issues in using ratings as performance criteria include (a) design of the rating form to be used, (b) advantages and disadvantages of ratings from different sources (e.g., supervisors or peers), (c) type of training to provide to raters, (d) examination of the performance judgment process to aid in understanding and improving ratings, and (e) evaluation of ratings in relation to psychometric properties (e.g., halo, reliability, validity, and accuracy). Because the fifth issue is relevant to each of the first four, evaluation of ratings will be discussed first, followed by discussions of the other four issues.

Evaluation of Ratings. The following subsections describe various approaches to evaluating job performance ratings.

Psychometric Properties. Ratings of job performance often suffer from psychometric errors. These can be classified as:

- *Distributional errors.* These errors involve raters misrepresenting the distributions of performance across persons they are evaluating. Misrepresentations can occur both in the means of ratings they provide (leniency/severity) and in the variance of ratings that result (restriction of range). Regarding leniency/severity, a rater may provide evaluations that are higher than warranted by actual performance levels (leniency) or lower (severity). This bias can be caused by raters having inaccurate frames of reference or norms that result in inflated or deflated ratings. With restriction of range, a rater may rate two or more ratees on a dimension such that the variance of these ratings is lower than the variance of the actual performance levels for those ratees. Raters who commit this error fail to differentiate sufficiently between ratees on individual dimensions.

- *Illusory halo.* A rater might make ratings on two or more dimensions such that the correlations between the dimensions are higher than between-dimension correlations of the actual relevant behaviors. This halo effect arises when the rater fails to differentiate sufficiently between performance *on different dimensions* for individual ratees. It is also possible that raters might underestimate the relationships between dimensions and provide ratings that correlate lower than the actual behaviors (Fisicaro, 1988).

- *Other errors.* Not as commonly referred to are such perceptual and rating errors as the similar-to-me error (Latham, Wexley, & Pursell, 1975), the first impression error (Latham et al., 1975), and error due to systematic distortion (Kozlowski & Kirsch, 1987), similar to the logical error of Guilford (1954). The *similar-to-me error* refers to an unwarranted projection of a rater's own personal characteristics onto a ratee. *First impression error* occurs when the rater allows early experiences with a ratee to be weighted more than they should be. *Systematic distortion* is an error characterized by a rater making evaluations of ratees on multiple dimensions such that the pattern of correlations between dimensions more closely reflects the semantic similarity of the dimension

names than the actual correlations between behaviors relevant to those dimensions. Systematic distortion is said to occur when the rater makes assessments based on assumptions about what behaviors *should* go together instead of according to the actual covariation of the behaviors (Shweder, 1975).

Interrater Reliability. Interrater agreement in performance evaluations *within* rating source (e.g., between peers) or *across* sources (e.g., between supervisors and peers) is sometimes offered as indirect evidence for the accuracy of ratings. From a logical perspective, this is of course a problem. Raters within *or* across sources may agree closely in their evaluations of ratees but may still be inaccurate. They may, for example, *all* focus on ratee characteristics (such as likability) that are irrelevant to job performance.

Campbell, Dunnette, Lawler, and Weick (1970) and Borman (1974) made the further point that high interrater reliability between raters at different organizational levels (e.g., peers versus supervisors) should not necessarily be expected. Members of different organizational levels may have different orientations and perspectives regarding ratee performance because of differences in roles related to the ratees, and they may observe significantly different samplings of ratee behavior as a result. Accordingly, raters from different levels may not agree very closely in their evaluations, but each may be providing valid performance data based on relevant but somewhat different performance information.

On balance, good interrater agreement is desirable, especially between raters at a single organizational level. Such agreement suggests that at least they are focusing on similar samples of job behavior, although, as mentioned, care must be taken in interpreting lower interrater reliabilities.

Convergent and Discriminant Validity. Kavanagh, MacKinney, and Wolins (1971) presented an ANOVA approach for evaluating the convergent and discriminant validity of ratings. A main effect and an interaction term in the ANOVA model are especially useful for these purposes. The ratee effect indexes convergent validity or overall interrater agreement collapsed across dimensions. The ratee x dimension interaction indicates the degree of discriminant validity, or the agreement between raters in the *patterns* of individual ratees' performance levels on the different dimensions. Kavanágh et al. described intraclass indices that can be used to compare levels of convergent and discriminant validity across different studies.

As Schmitt and Stults (1986) point out, a limitation of the ANOVA approach is that trait intercorrelations and method intercorrelations cannot be estimated. More recent developments, using path analysis (e.g., Althauser, 1974) or confirmatory factor analysis (e.g., Kenny, 1979; Widaman, 1985), overcome this limitation and, further, allow evaluation of convergent and discriminant validity hypotheses at both the matrix and individual measure levels.

The main thrust of the Widaman approach, for example, is to compare the fit of different hierarchically nested models to estimate in a multitrait-multimethod matrix of ratings (a) convergent validity (a model with dimension or trait factors fits better than one with no such factors present); (b) discriminant validity (a model with two or more interpretable trait factors fits better than a model with a single trait factor); and (c) method variance—that is, different rating instruments or rating sources (a model with method factors fits better than one with no method factors). A distinct advantage of this strategy is that variance-accounted-for estimates can be computed to index precisely the extent of convergent validity and method bias in the ratings. The general approach has been described and applied to

performance rating data by Vance, MacCallum, Coovert, and Hedge (1988).

Accuracy. Some have argued that *accuracy* is the most important criterion for evaluating the quality of performance ratings (e.g., Bernardin & Pence, 1980; Borman, 1977; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982).[1] Investigating psychometric properties, interrater agreement, and convergent/discriminant validity of ratings certainly yields useful information about such measures; but these are only indirect methods of assessing accuracy. The argument is similar to preferring information about a test's validity to such qualities as its reliability or item characteristics. The correct rank ordering of ratees on a performance dimension and/or on overall performance is, for example, needed to ensure that test validities are properly estimated. In fact, for any application in personnel research that requires evaluation of performance at the individual level, validity or accuracy of criterion measurement is very important.

It should be mentioned that this requirement for validity or accuracy of performance scores in personnel research is somewhat in contrast to what is needed for performance appraisal systems intended to serve administrative purposes such as promotion decisions and employee feedback and development. In those cases, evaluative criteria such as acceptance by users (e.g., Banks & Murphy, 1985) and usefulness in meeting organizational and individual objectives (DeVries et al., 1981) become equally important. As a concrete example, when evaluating a sensitive employee who is performing poorly but has excellent potential for effective performance and seems to need some encouragement to realize that potential, accuracy in rating the poor performance seems less important than providing positive feedback on some of the tasks performed well, thus enhancing the employee's future performance. More generally, when an administrative performance appraisal system is not accepted as fair and useful by employees, accuracy of the ratings made with the system appears to be less important than gaining system credibility. However, in the case of performance rating for personnel research applications, the requirement for accuracy (or validity) is more evident.

Unfortunately, evaluation of accuracy or validity requires external criterion target scores of some type against which to compare the ratings. Such scores are almost never available in organizational environments. Therefore, accuracy research has largely proceeded in laboratory settings.

To evaluate accuracy, written or videotaped vignettes have been developed and target performance scores for the vignettes have been derived. With the "paper-people" approach to vignette development (e.g., DeCotiis, 1977), behavioral examples scaled according to effectiveness level are woven together to create descriptions of hypothetical employees performing on jobs. Typically included in each vignette is one behavioral example representing each performance dimension for the job, selected to reflect the intended performance level. Ratings of the hypothetical employees' performance can then be compared dimension by dimension to the effectiveness scale scores of the behavioral examples in the vignettes.

Videotaped vignettes of employees performing tasks or jobs can also be developed to help investigate accuracy in ratings (Borman, 1977; Murphy et al., 1982). Target scores are typically assigned to videotaped performances on each dimension, using "experts" thoroughly familiar with the task or job to provide performance judgments. The procedure involves giving the expert judges considerable exposure to the taped behavior so their ratings are as well informed as possible. High convergent and discriminant validity across judges is usually offered as evidence for the quality of the means of the experts' target scores (e.g., Borman, 1977;

Murphy et al., 1982). Once target scores are assigned to the paper-people or videotaped vignettes, these stimulus materials can be employed in rating research, and the accuracy of performance ratings can be assessed, as appropriate.

Regarding actual measures of accuracy, many years ago Cronbach (1955) greatly clarified the issue of indexing interpersonal accuracy by demonstrating that the often-used difference score measure (*D:* Sum of the differences between ratings of persons on two or more traits or dimensions and the target criterion scores for those same persons and dimensions) decomposes into four separate components of accuracy, each with a different psychological interpretation. He argued against use of the D or D^2 measures of accuracy because they confound the four components. The first component, *elevation*, refers to how closely a rater's grand mean of his or her ratings (across ratees and dimensions) agrees with the grand mean of the target scores. *Differential elevation* relates to accuracy in differentiating between different ratees' mean target scores (collapsed across dimensions). *Stereotype accuracy* reflects how correctly raters differentiate between dimension target score means (collapsed across ratees). Finally, *differential accuracy* is the degree to which raters correctly rank order ratees on each dimension, controlling for ratee and dimension effects (see Cronbach, 1955, or Murphy et al., 1982, for the actual mathematical definitions of the four accuracy components).

Differential elevation *(DE)* and differential accuracy *(DA)* seem especially useful for applications in performance rating research, *DE* because it indexes a rater's accuracy in assessing different ratees' overall performance and *DA* because it provides an index of skill in differentiating between different ratees' performance levels in individual aspects of the job. Stereotype accuracy might be useful in assessing a group's training needs by accurately describing the group's strong and weak performance areas averaged across its members.

In addition to the D or D^2 measure and Cronbach's component scores, other accuracy or validity indices have been used. Borman (1975) and Athey and McIntyre (1987) correlated ratings with criterion target scores on several dimensions *by individual ratees* and then averaged these correlations across the ratees evaluated. This is a validity index of how correctly raters can identify strong and weak performance areas for each ratee, and may be especially appropriate for evaluating interpersonal diagnostic skill in performance counseling and management development settings. Gordon (1970) and Thornton and Zorich (1980) have studied what might be called observational accuracy, or success in recognizing and recalling behaviors that have occurred in performance.

Still another way to study accuracy is to use signal detection theory (Baker & Schuck, 1975; Lord, 1985). Lord, for example, argued that estimating hit rates (proportion of observed items correctly identified as exhibited by the target ratee) and false-alarm rates (proportion of items falsely so identified) provides data that allow inferences to be made about the accuracy of the information processing strategies used by individual observers. The basic notion is that raters may be quite accurate using a Cronbach index but still be inaccurate in the ratee behavior they actually process in arriving at their ratings. In this sense, signal detection theory provides a more precise, fine-grained assessment of rating accuracy.

In sum, there are two major issues in studying accuracy in ratings. The first is establishing criterion target performance scores against which to compare ratees' ratings. This is an extremely difficult problem. Target criterion scores are simply very hard to justify in settings that are at all realistic. The most realistic rating task designed to include target scores would seem to involve the videotaped performances of individuals on jobs, but even this task is obviously very different from the typical performance rating task faced by persons in

organizations (even from a for-research-only rating task). To increase the realism of the rating task, Feldman (1981) suggested giving subject raters in videotape research other responsibilities besides viewing and rating ratees. Favero and Ilgen (1989) implemented this suggestion by requiring raters in a videotape lab study to perform other duties in addition to evaluating the videotaped performers (e.g., working on organizational problems presented in an in-basket). Further, Bernardin and Villanova (1986) described a "modal rating situation" typically faced by persons in organizations actually responsible for performance appraisal. More attention must be paid to realism and generalizability issues.

The second issue is how to compute accuracy. This is less problematic. Cronbach's (1955) early insights in the field of interpersonal perception have provided good choices of accuracy indices. The researcher need only select an index or indices conceptually appropriate for the research being conducted.

Relations Between Accuracy and Rating Errors. Intuitively, we would think psychometric errors in ratings should be correlated negatively with accuracy. The greater the error in a set of ratings, the less accuracy to be expected. If this relationship is high, we might even employ indices of psychometric error as substitute measures of (in)accuracy. This would be very useful because, as we have seen, accuracy is difficult to study.

However, two kinds of research results suggest that reducing psychometric error in ratings may not enhance their accuracy. In studies where both psychometric errors *and* accuracy scores could be computed for individual raters, correlations between accuracy and rating errors such as leniency, halo, and restriction of range are near zero (Murphy & Balzer, 1989). Low *positive* correlations between halo error and accuracy have actually been reported (Cooper, 1981). In the other type of study, when raters are trained successfully to reduce psychometric error, accuracy is either not affected (Borman, 1979a) or actually decreases (Bernardin & Pence, 1980).

Recent work has shed more light on rater error–accuracy relationships. Becker and Cardy (1986) demonstrated that relationships between halo and accuracy vary according to the particular measures of halo and accuracy used, and counterintuitive relationships between these measures (i.e., positive correlations between halo and accuracy) can be partially explained by the way the halo and accuracy indices are defined statistically. Fisicaro (1988) found that when halo was defined as an absolute difference between dimension interrelations of ratings and dimension interrelations of actual target performance scores (i.e., taking into account "negative halo" as well as "positive halo"), correlations between this conception of halo error and accuracy were more negative—as one would intuitively expect. Regarding the halo measures, note that the variance-across-dimensions, within-ratee index of halo was found to be flawed in that, unlike the mean intercorrelations of ratings on the dimensions, it is influenced by characteristics of the ratings irrelevant to halo (Pulakos, Schmitt, & Ostroff, 1986). In any case, although motivation is considerable for using estimates of psychometric error as proxy measures of accuracy, relationships between these rating errors and accuracy are not sufficiently well specified for this approach to be feasible.

Often, research on performance ratings has properly focused on ways to improve their usefulness as criteria. This effort has proceeded in four principal areas of research: (a) rating format design and empirical comparisons between formats, (b) selection of raters, (c) rater training, and (d) rating process research. Each will be discussed in turn.

Research on Rating Formats. Over the years, many different types of rating formats have been developed and used in research and practice. Table 1 lists many of these formats. Refer

TABLE 1

Performance Rating Formats

- Graphic scales (Paterson, 1922–23)
- Man-to-man scales (Guilford, 1954)
- Ranking forms (Ghiselli & Brown, 1955)
- Forced choice scales (Bartlett, 1983; Sisson, 1948)
- Summated scales (J. P. Campbell, Dunnette, Arvey, & Hellervik, 1973)
- Critical incidents checklist (Flanagan, 1954)
- Behaviorally anchored rating scales—BARS (P. C. Smith & Kendall, 1963)
- Behavior observation scales—BOS (Latham & Wexley, 1981)
- Behavior summary scales (Borman, 1979a)
- Mixed standard scales (Blanz & Ghiselli, 1972)
- Behavioral checklist (Komaki, 1981)
- Frequency of behavior scale (Kane, 1986)

to Bernardin and Beatty (1984) and Whisler and Harper (1962) for descriptions of most rating formats.

It has seemed compelling to believe that characteristics of rating formats are important determinants of rating accuracy. There have been certain very creative and conceptually sound ideas about format development. Here are some highlights:

- The notion of supervisors or peers providing numerical scores for employees on job-relevant traits or performance areas is an interesting idea. Ideally, it provides well-informed observers with a means of quantifying their perceptions of individuals' job performance. This is highly preferable to verbal descriptions of performances because individuals can now be compared in a reasonably straightforward way. The notion can be viewed as analogous to developing structured job analysis questionnaires to take the place of verbal job descriptions for purposes of comparing jobs (McCormick, 1976; see also Harvey, this volume). In each case, quantification of perceptions clears the way for scientific study of an area that could not previously be studied in this manner.
- Development of forced choice scales was an ingenious attempt to overcome problems with raters' subjectivity and bias in making performance evaluations. The main idea was to eliminate the *opportunity* for raters to slant ratings according to their own subjective biases. One version of the scales developed to evaluate Army officers presented descriptive behavioral items in groupings of four (Sisson, 1948). Two relatively favorable, positive items were matched in rated social desirability, but one of the items was judged very descriptive of effective performers and the other judged less so. Likewise, there appeared two items with relatively low (and equal) rated social desirability, but one was judged substantially more descriptive of poor performers than the other. An example tetrad grouping is presented as follows:

	Most Descriptive	*Least Descriptive*
A. Cannot assume responsibility	❑	❑
B. Knows how and when to delegate authority	❑	❑
C. Offers suggestions	❑	❑
D. Changes ideas too easily	❑	❑

The rater was asked to review the items in each grouping and to check which of the

four items was most descriptive and which was least descriptive of the ratee. A score of +1 was given for responding "most descriptive" to the positively keyed item or "least descriptive" to the negatively keyed item, and a score of –1 was given for responding "least descriptive" to the positively keyed item or "most descriptive" to the negatively keyed item. Responding either most or least descriptive to the nonkeyed items earned a score of zero.

Proponents of this kind of scale argue that the hidden-key feature of the scale design prevents raters from assigning higher (or lower) ratings than warranted. One criticism of the format is that a single overall performance score is obtained from the ratings rather than a score for each different performance dimension [although King, Hunter, & Schmidt (1980) constructed multidimensional forced choice scales]. A second problem is a consequence of the scale's main advantage: Raters have expressed dissatisfaction about not having control over the outcomes of their ratings. Nonetheless, this format represents a bold initiative to create a relatively objective rating instrument.

- P. C. Smith and Kendall (1963) extended the notion of critical incidents (Flanagan, 1954) by designing a rating format they referred to as behavioral expectation scales, now generally labeled *behaviorally anchored rating scales* (BARS). P. C. Smith and Kendall reasoned that different effectiveness levels on job performance rating scales might be anchored using behavioral examples of incumbent performance. Accordingly, they developed performance rating dimensions with scaled behavioral examples anchoring the appropriate effectiveness levels on the dimensions. The high and low segments of a rating dimension with two behavioral anchors for the safety-mindedness dimension on the job of power plant maintenance worker (Bosshardt, Rosse, & Peterson, 1984) appear in the example that follows:

7 ▪ This employee stopped another employee from "air lancing" coal dust from a pulverizer without wearing proper eye protection. As a result, possible eye damage was prevented.

1 ▪ This employee was replacing a pipe union on an acid line without wearing any protective equipment. The union broke with pressure on the line, and his face received an acid burn.

Essentially, the rater's task is to compare observed job behaviors of the ratee with the behavioral anchors on the scale to assign a rating on that dimension. This was seen as preferable to evaluating a ratee without guidance regarding the effectiveness levels of different scale points. The BARS idea is more than a format; it is a system, or even a philosophy (Bernardin & Smith, 1981). For example, ideally raters should record examples of employee work behavior

throughout the appraisal period to aid in assigning performance ratings.

Another positive feature of BARS is that users of the system typically participate in scale development, enhancing the credibility of the format. Further, from a domain sampling perspective, BARS development steps provide an excellent vehicle with which to identify all important performance dimensions for a job (J. P. Campbell, Dunnette, Arvey, & Hellervik, 1973). Having persons knowledgeable about a job generate many actual behavioral examples of performance on that job should result in an exhaustive listing of its performance requirements and provide an operational way to define comprehensively the *conceptual criterion*, in Astin's (1964) nomenclature.

- Blanz and Ghiselli (1972) introduced the mixed standard scale (MSS), a rating format with several appealing features. The MSS consists of three behavioral statements, essentially BARS anchors, for each performance dimension. One reflects relatively effective performance, a second represents midlevel or average performance, and a third depicts lower-level performance. Typically, the behavioral statements across dimensions and effectiveness levels are randomly ordered on the scale, and the rater is asked to indicate whether the ratee's performance is worse than, the same as, or better than the performance represented in each statement. A score for a ratee on each dimension can then be derived according to the following rules for these logically consistent rating combinations, where a plus sign means "better than" ratings, zero means "the same as" ratings, and a minus sign means "worse than" ratings:

Effective Statement	*Average Statement*	*Ineffective Statement*	*Derived Rating*
+	+	+	7
0	+	+	6
–	+	+	5
–	0	+	4
–	–	+	3
–	–	0	2
–	–	–	1

Actually, there are 27 possible patterns of ratings for a dimension; the other 20 represent illogical combinations. For example, 0, –, + is not a logical set of ratings. How can a ratee be at the same level of performance as the high effectiveness statement but perform worse than the midlevel statement? This feature of the MSS can be viewed as a distinct advantage, because the number of illogical response patterns can be scored for each rater, ratee, and dimension, and useful inferences can be made from these scores. For example, if such scores are high for a rater they may indicate incompetently completed ratings by that rater; high scores for a ratee may suggest that the ratee is difficult to evaluate; and high scores for a dimension may mean that the statements associated with the dimension are ambiguous. This diagnostic information can then lead to useful interventions, such as rater training for inconsistent raters, a search for alternate raters for ratees scored inconsistently, and further scale development work for high error dimensions. Another potential advantage of the MSS format with behavioral statements randomly ordered instead of grouped by dimension is that halo might be reduced because the dimensions are disguised from raters, although a recent study found no decrease in halo as a function of this feature of the MSS format (Dickinson & Glebocki, in press).

This format thus represents an unusual and potentially effective way to generate

performance information. The judgments a rater is asked to make are somewhat more straightforward than is the case with BARS, for example. The rater must simply compare the effectiveness of observed ratee behavior to the effectiveness reflected in a single behavioral statement, without reference to dimensions of performance or continua of effectiveness.

Format Comparison Studies. A reasonable empirical question concerning these and other rating formats is, Which ones are better? Format comparison studies have been conducted to address this issue, usually employing psychometric criteria as the dependent variables. Early studies (e.g., Blumberg, DeSoto, & Kuethe, 1966; Madden & Bourdon, 1964; Taylor, Parker, & Ford, 1959) focused on relatively narrow considerations—number of scale points, vertical versus horizontal scales, and the like. More recent work compared entire formats and used such criteria as halo, leniency, restriction of range, and interrater reliability. As an example of such a study, Bernardin (1977) had college student raters evaluate their college professors on BARS and two carefully developed summated rating scales. The latter had positively and negatively worded behavioral statements and required raters to indicate how frequently (from never to always) each ratee exhibited them. Results showed no significant differences between formats on the psychometric properties of interrater reliability, halo, leniency, and discrimination between ratees.

Reviews of format comparison studies (e.g., Landy & Farr, 1980; Schwab, Heneman, & DeCotiis, 1975) suggest the following conclusions: The psychometric superiority of BARS is questionable. Some studies show ratings on BARS have better psychometric properties than ratings made on other formats (e.g., J. P. Campbell, Dunnette, Arvey, & Hellervik, 1973), but other studies show no such differences (e.g., Bernardin, Alvares, & Cranny, 1976). The most important consideration in format development may be that rigorous scale development procedures are followed (Bernardin, 1977). Landy and Farr (1980) estimate that as little as 4 percent of the variance in psychometric quality may be accounted for by format, although Guion and Gibson (1988) argue that it is premature to give up on format-related research.

A useful way to view performance rating formats is as mechanisms for helping raters (a) conduct an organized and efficient search for ratee performance-related behavior, (b) translate these behavioral observations into evidence pertinent to assessing ratee performance on each dimension, and then (c) make accurate judgments about ratee effectiveness levels on each dimension. Ideally, formats should be configured so that the operations required of raters reflect natural cognitive processes leading to efficient and effective processing of performance information. Characteristics of rating scales should provide a clear presentation of standards to help raters evaluate observed work behavior. Thus, it should be useful to identify and then use those features most compatible with effective observation and evaluation of behaviors.

Feldman (1986) explores these fundamental rating format characteristics in an essay in which he distinguishes between performance measurement systems with dimensions permitting *analytic* processing and those with dimensions requiring *intuitive* processing. In the analytic case, performance on the dimensions can be objectively defined, with relatively few alternatives to a specified way of performing the job on those dimensions. Feldman (1986) offers the example of the machinist job, where the frequency and magnitude of errors largely define performance on the technical skill dimensions. In this case, a format might be best designed to require recall or estimation of the frequency, rate, or intensity of

behaviors that job analysis indicates are important for successful performance.

In the intuitive case, performance on the job must be assessed according to dimensions based on value systems not as objectively specifiable as in the analytic case. Managerial jobs represent a good example of where intuitive assessment of performance is required. The role of the rater is then to interpret ratee behavior according to these values as they are depicted on the rating instrument. Thus, in such a case, formats should be designed to define in observable, behavioral terms the performance dimensions highly valued by the organization.

In sum, minor rating format manipulations are not likely to make much difference in improving the accuracy of performance ratings. However, in the design of rating instruments, it is useful to consider how best to depict a job's performance requirements on the instrument in a way that is easily understandable by raters as well as cognitively compatible with the rating task to help raters make accurate performance judgments.

Selection of Raters. A second attempt to improve the psychometric properties and accuracy of ratings concerns attention paid to the source of those ratings (e.g., supervisor, peer, or self). Logically, each of these and other sources has advantages in providing valid performance information. Experienced supervisors have reasonably good norms for performance, because typically they have seen relatively large numbers of employees working on the job and thus have well-calibrated views of different performance levels. Peers are usually privy to the most performance information regarding their fellow workers; lay wisdom suggests that it is difficult to hide your actual performance level from co-workers. Self-ratings have a similar advantage in that, clearly, considerable performance-related information should be available from these ratings.

Other rating sources are used less often, but have certain inherent advantages. Supervisees are likely to have especially relevant information about their supervisors' leadership skills. An outside observer sampling on-the-job behavior is free from the possible biasing effects of organizational and personal roles related to the ratee.

There are disadvantages to each of these rating sources as well. Supervisors may not actually observe much of the day-to-day work performance of supervisees. Co-workers and supervisees often lack experience in making formal performance evaluations, and the latter are typically in a position to see only a relatively small portion of their supervisors' job performance. Self-ratings may be distorted due to inflated evaluations of the rater's own performance. Finally, an observer will not usually view a sufficient sample of ratee performance to obtain an accurate picture of typical performance over time. The previous statements depend to some extent on the structure of the ratee's organization and the particular interactive work roles practiced by supervisors, co-workers, and supervisees in the organization.

Research on this topic seeks to determine which source or configuration of sources provides the most accurate, error-free evaluations of performance. Accordingly, studies have been conducted to (a) assess the psychometric properties and reliability of ratings from each source, (b) examine the interrater agreement between sources, and (c) evaluate the predictive validity of peer and self-assessments.

Regarding the first type of study, conclusions are that self-ratings are generally more lenient than peer or supervisor ratings (e.g., Kirchner, 1965; Parker, Taylor, Barrett, & Martens, 1959) but contain less halo than do ratings from those sources (e.g., Heneman, 1974; Kirchner, 1965). Results are mixed regarding comparative levels of interrater reliability within the peer and supervisor sources, but on balance, supervisory ratings tend to be more reliable (Klieger & Mosel, 1953; Pulakos & Borman, 1988; Springer, 1953).

In relation to interrater agreement between sources, peer and supervisor ratings typically agree more closely with ratings from either of the other two sources than do self-ratings (Harris & Schaubroeck, 1988; Klimoski & London, 1974). Yet interrater reliability between peers and supervisors is usually only moderate, with agreement within rating source greater than agreement across the two sources (e.g., Berry, Nelson, & McNally, 1966; Borman, 1974; Gunderson & Nelson, 1966).

Although high interrater reliability is desirable in ratings, suggesting that different raters are focusing on similar, presumably job-related factors, close agreement across rating sources may be unrealistic, as I mentioned earlier. Thus, low to moderate across-source agreement in ratings may not be so much a sign of unreliability as an indicator that somewhat different aspects of performance are being observed and reported on (Borman, 1974; J. P. Campbell et al., 1970).

Finally, peer and self-assessments have been used to predict future performance. This can be conceived of as a kind of validity check on ratings from these sources. Among others, Hollander (1954, 1965), Downey, Medland, and Yates (1976), and Waters and Waters (1970) have evaluated relationships between peer assessments and later performance in both military and civilian samples. Kane and Lawler (1978) summarized this work and found a mean validity coefficient of .43 when peer nominations (i.e., identification of the best and worst performers) were used to predict subsequent performance and a somewhat lower validity when peer ratings were used. These results are impressive in a sense, but interpreting the magnitude of the correlations as indicators of validity is problematic. They may underestimate validity where the constructs reflected in the peer assessments and subsequent criteria are dissimilar (e.g., ratings of present performance on technical proficiency against a criterion of supervisory performance). They may overestimate validity if the criteria for the ratings are also ratings and the predictors and criteria share invalid method variance. Nonetheless, there is evidence that peer assessments are tapping important performance-related variance (Kane & Lawler, 1978).

Mabe and West (1982) performed a meta-analysis of relationships between self-assessments on traits or competency dimensions and criteria relevant to those assessments, including objective test scores, academic grades, and supervisor ratings. The mean correlation was .29, indicating moderate validity for self-ratings.

On balance, however, peer and supervisor ratings of performance appear to hold the most hope for providing accurate depictions of job performance. Leniency in self-ratings and (perhaps worse yet) differential leniency, in which some raters inflate their self-ratings more than others, are especially problematic with self-assessment.

Rater Training. Rater training provides a promising approach to improving the quality of performance ratings. Two general kinds of training programs have emerged to help raters generate more error free and accurate ratings (Bernardin & Buckley, 1981; D. E. Smith, 1986). *Rater error training* seeks simply to alert raters to certain psychometric or perceptual errors such as leniency/severity, halo, restriction-in-range, and similar-to-me effects. Training often takes the form of a brief lecture on or demonstration of the error and a plea to avoid such errors when making performance ratings (Bernardin & Buckley, 1981). *Frame-of-reference training* (Bernardin & Pence, 1980) attempts to convey to raters that performance is multidimensional and to familiarize them thoroughly with the actual content of each performance dimension. Regarding familiarization, examples of different levels of performance on individual dimensions are typically reviewed with raters, along with the "correct" or actual performance levels the examples reflect (e.g., Pulakos, 1984).

Researchers have conducted studies comparing the psychometric properties and accuracy of ratings made by raters trained using one of the approaches just discussed and ratings generated by untrained raters. Results suggest the following conclusions:

- Error training is usually successful in reducing the target psychometric error (e.g., Latham, Wexley, & Pursell, 1975).
- Error training does not improve the quality of ratings when interrater reliability or accuracy is used as a criterion (e.g., Borman, 1979a).
- Frame-of-reference training increases rating accuracy (McIntyre, Smith, & Hassett, 1984; Pulakos, 1984).

In addition, practice in making ratings and feedback about rating errors and accuracy are important components of training programs (Latham, 1986; D. E. Smith, 1986). The five frame-of-reference studies identified by D. E. Smith (1986) all employed practice and feedback in their successful programs.

A useful observation has been offered by Bernardin and Pence (1980): Rater error training is successful in reducing the target psychometric response set or error (e.g., halo), but essentially *new response sets* are forced on raters (e.g., to eliminate halo, spread out your ratings across dimensions), resulting in either no change in accuracy or a reduction in it. Similarly, Borman (1979a) suggested that to direct persons to adjust their rating distributions in some manner is relatively easy for training to accomplish; it is much more difficult to train raters to be more accurate. Frame-of-reference training appears to be the best bet to attain this worthwhile goal.

Research is needed now to identify the elements responsible for rater training success (e.g., Athey & McIntyre, 1987). This will allow streamlining and refinements of frame-of-reference programs to enhance their efficiency and effectiveness in improving rating accuracy.

Rating Process Research. Arguably, the so-called process approach to studying performance ratings began with Wherry's treatise on ratings (appearing in Landy & Farr, 1983, and Wherry & Bartlett, 1982). This ambitious effort presented a "theory of ratings," which included a kind of "job analysis of the rating process" (Landy & Farr, 1983, p. 285). Using applied measurement theory along with principles from the areas of human learning and memory, Wherry fashioned a series of propositions to help better understand this process.

Regarding attempts to illuminate the performance rating process, the basic rationale has been that we should "get beyond" manipulations of rating formats and other psychometric concerns with ratings to study in detail the entire sequence that raters follow in making performance judgments. Rating process research could then inform us enough to be able to intervene in order to reduce rater errors, biases, and inaccuracies. The reasoning goes that a scientific examination of performance rating requires a rigorous sequencing of theory development, hypothesis or proposition generation, and hypothesis testing. Cognitive psychology has been the source of models relating to the now familiar steps of observing, encoding, storing in memory, retrieving from memory, judgment, and rating (e.g., De Nisi, Cafferty, & Meglino, 1984; Landy & Farr, 1980). Personality and social psychology contributed additional concepts, such as implicit personality theory, attribution theory, and personal construct theory, to be considered when further conceptualizing and studying the performance rating process (e.g., Borman, 1983; Feldman, 1981; Ilgen & Feldman, 1983). These models and concepts will be reviewed next; then research applications will be discussed.

Rating Process Models. Two kinds of rating process cognitive models with somewhat different emphases have emerged in the performance rating literature. The first are termed *process models,* depicting the rating sequence presented in the previous paragraph, along with factors hypothesized to influence that process (Cooper, 1981; DeCotiis & Petit, 1978; De Nisi et al., 1984; Landy & Farr, 1980, 1983). The second kind of model emphasizes and elaborates on the encoding step in that sequence to consider in some depth categorization in processing performance-related information (Feldman, 1981; Ilgen & Feldman, 1983; Lord, 1985).

To provide a more detailed view of these two types of models, prototypes of each will be described. Then, the potential influences of implicit personality theory, attribution processes, and personal construct theory on performance judgments will be discussed briefly.

The De Nisi et al. (1984) model is comparatively detailed in specifying the cognitive steps that presumably take place during the rating process (see Figure 3). Performance information is sought and encoded and stored first in "individual memory bins" and then in longer-term memory. Before a performance evaluation is made, the rater makes judgments about possible external influences on the performance and how typical this performance is of the ratee. De Nisi et al. emphasize the rater as an active seeker of performance information; they also note the central importance of memory in the rating process. Thus, De Nisi et al.'s model contains the basic steps of observing, encoding, storing, retrieving, judging, and rating, as presented in several additional models (Cooper, 1981; Landy & Farr, 1980, 1983).

Each of these other process models has unique and useful features as well. For example, Cooper's formulation carefully attends to how different steps in his sequential process model influence both accuracy and illusory halo in ratings. Essentially, random error in the process sequence decreases accuracy, and systematic error in the sequence increases halo. Landy and Farr view performance rating as dependent upon "highly filtered information." The actual ratee behavior observed is influenced by several factors before emerging as a performance judgment about the ratee.

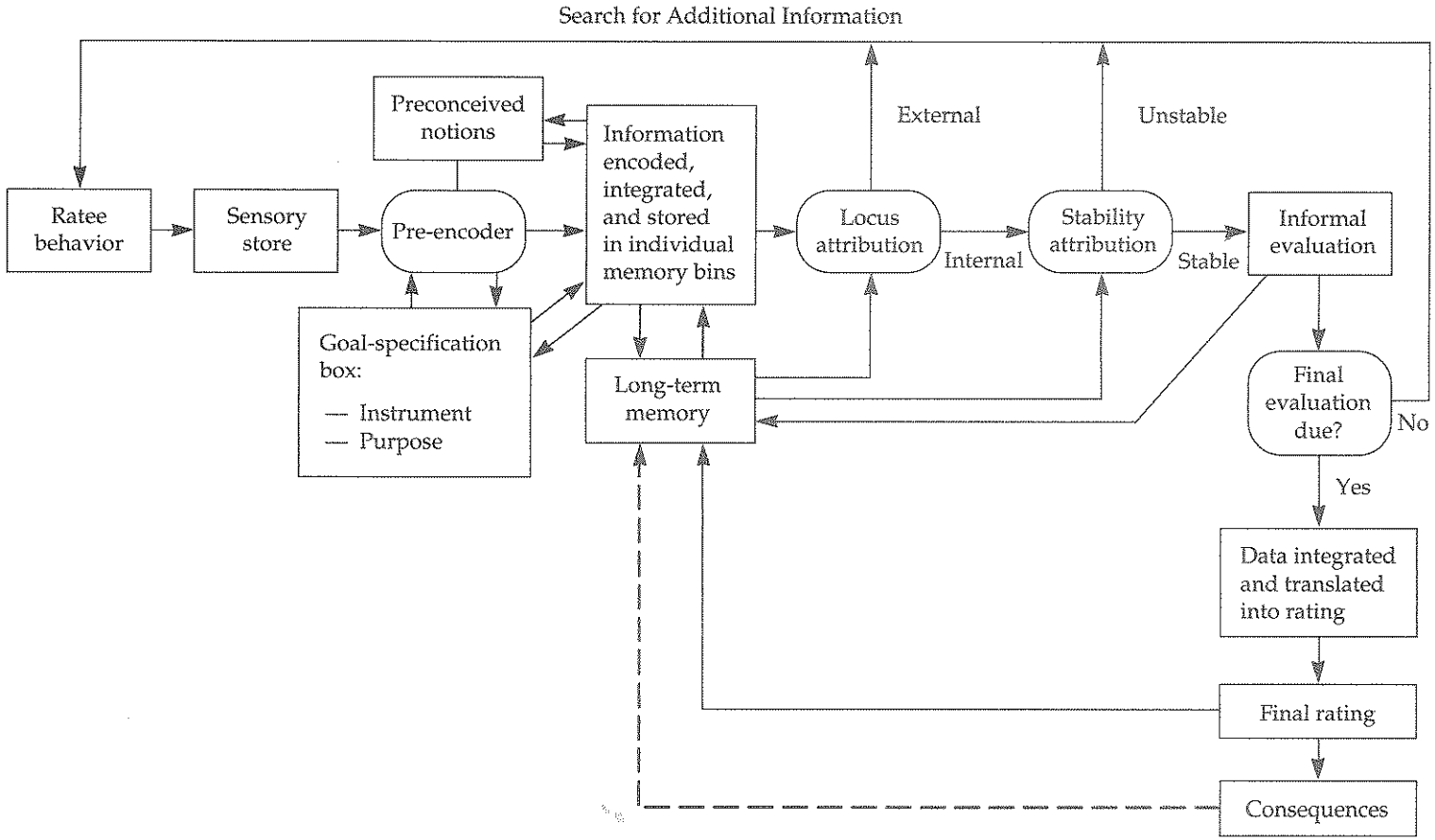
Feldman (1981) and Ilgen and Feldman (1983) provide prototypes of a somewhat different kind of rating process model. Although the cognitive-based information processing sequence described previously is incorporated into their models, two additional features are emphasized. First, these authors elaborate considerably on categorization processes, referring to that part of the process model where encoding is taking place. Confronted with a barrage of performance-related information about ratees, the rater simplifies the information by categorizing it into dimensions that represent in relatively simple form the complexity of the "raw" behavior observed. Categories are selected for a ratee behavior via a matching process between features of the behavior and the category (e.g., hard-working, slacking off), and when work-related information about the ratee is to be recalled, often the category is brought up rather than the specific behavior.

A second difference between this type of model and those discussed previously is that automatic and controlled attentional processes are distinguished. These authors make the point that when the patterns of ratee behavior conform with previous impressions, then that behavior is "automatically" categorized without much conscious effort. However, when an unexpected or otherwise noteworthy behavior is observed, more active categorizing, including changing categories for a ratee (e.g., from conscientious to careless at times), is likely to occur.

Categorizing performance-related behavior to simplify the large amount of performance information observed is an important process

FIGURE 3

De Nisi, Cafferty, and Meglino (1984) Model of Performance Rating



From "A Cognitive View of the Appraisal Process: A Model and Research Propositions" from *Organizational Behavior and Human Performance* by A. S. De Nisi, T. P. Cafferty, and B. M. Meglino, 1984, *Organizational Behavior and Human Performance, 33*, p. 363. Copyright 1984 by Academic Press, Inc. Adapted by permission.

to understand. Research in cognitive psychology has confirmed the heuristic usefulness of some kinds of knowledge structures. Besides categories (similar to *dimensions* in performance rating parlance) as aids in this simplification process, schemata, prototypes, stereotypes, and scripts have been discussed as important in social perception.

Briefly, schema is a generic term that subsumes several other hypothesized cognitive structures. *Schemata* are virtually synonymous with categories, both referring to reference concepts used by raters to help make judgments about other persons (e.g., Cantor & Mischel, 1977; Wyer & Srull, 1980). *Prototypes* highlight modal or typical features of a category (e.g., Hastie, 1981) and can be thought of as good examples of schemata. An examplar of a prototype is, "Al is a perfect example of what I think of as sociable." *Stereotypes* are similar to prototypes but refer to groups of people rather than individuals (Hamilton & Gifford, 1976). In addition, stereotypes tend to carry a significant affective component, usually negative. Finally, *scripts* are events or event sequences that are remembered as being representative of a person's actions (Abelson, 1981). They are often abstracted versions of actual events, with gaps filled in to create a coherent story. In filling these gaps, actions and other made-up parts of the story are included to be consistent with what is remembered about the event sequence related to the person being evaluated.

Thus, basically, schemata and associated hypothesized knowledge structures are used to reduce complexity in social perception. Loss of specific behavioral detail may lead to errors and biases in perception and judgment.

Relevant Personality/Social Psychology Concepts. Person perception concepts from the areas of personality and social psychology have also been useful in contributing to thought and research on the performance judgment pro cess. Implicit personality theory (IPT), personal construct psychology (PCP), and attribution processes have provided alternative framworks from which to view performance ratings.

Marked similarities are evident between certain features of IPT and PCP and the concept of a schema. IPTs (Schneider, Hastorf, & Ellsworth, 1979) have to do with assumptions a person makes about how personal characteristics (or work behaviors) covary in people, whether these assumptions are right or wrong. The personal characteristics in IPTs are then similar to schemata, but the *relationships between* different schemata are the focus here. Some research shows that relationships between personal characteristics based on ratings of others well known to the raters are quite similar to assumed relationships between these personal characteristics, suggesting in turn that ratings are made on the basis of *assumed* relationships between personal characteristics rather than according to *actual* relationships (Hakel, 1974; Passini & Norman, 1966). Another possible effect IPT may have on perceptions and ratings of personal characteristics (or performance-related behavior) involves individual differences between different raters' IPTs. Such differences in assumed correlations between dimensions could contribute to interrater disagreement.

Personal constructs (Adams-Webber, 1979; Kelly, 1955) are defined as content categories used to organize and simplify information. In particular, as part of his ambitious psychological theory, Kelly (1955) observed that individuals develop personal construct systems to judge events (or the activities of other people) and to make predictions about future events. Importantly, some of these categories are imposed on their person perceptions. These interpersonal filters may influence observations and judgments about other people by providing frames of reference or sets that make perceivers look for selective kinds of interpersonal information and interpret this information according to their own constructs (Duck, 1982). Accordingly, in the domain of social perception,

personal constructs are very similar to schemata.

The social cognition literature (e.g., Ostrom, Pryor, & Simpson, 1980; Wyer & Srull, 1986) is compelling in arguing for the existence in some form of these knowledge structures, schemata, IPTs, or personal constructs. However, the question might be asked, "How do these categories function in the performance evaluation setting?" How can the heuristic notions discussed in this literature be put into practice to determine more clearly the importance of these notions for influencing performance judgments? One possibility is to consider what might be referred to as "folk theories" of job performance (Borman, 1983). *Folk theories* are performance constructs used naturally by persons familiar with a job to describe its performance requirements and to differentiate between effective and ineffective performers. Two examples from job analysis interviews are (a) a secretarial supervisor stating that a key to effectiveness in his or her secretary's job is "maximizing time on task, staying with work tasks until they are completed," and (b) a sales manager reporting that a critical factor to successful performance in sales positions within the district is "knowing the products inside out." These firm opinions about job performance requirements, or folk theories, may be examples of categories or schemata that influence the ways organization members view and interpret work behavior of persons performing on the job. Accordingly, these categories or schemata could affect performance ratings made by supervisors or peers.

To test the importance of considering schemata in studying performance ratings, it is crucial to investigate relationships between the content and/or structure of categories and actual rating behavior. For example, do raters with very different schemata regarding the performance requirements for a job tend to disagree in their performance ratings? A related consideration is the stability of these categories and their structure over time and in different work contexts. Are categories difficult to change? Can "valid" categories from a job analysis be trained so that raters possess an effective category system?

Attribution theory is also relevant to our concerns about the performance rating process. *Attribution* refers to observers or raters assigning *causes* to behavior (Kelley, 1967). Specifically, the fundamental attribution error (Ross, 1977) occurs when individuals interpret their own behavior as caused primarily by situational factors, yet interpret behavior of others as influenced more by their personal characteristics or internal dispositional factors. This effect has been demonstrated in many studies (cf. Kelley & Michela, 1980).

Results from attribution research most germane to performance rating are, first, that consistent behavior (performance) is more likely to be attributed to dispositional factors than is inconsistent behavior (Frieze & Weiner, 1971). Second, and related to this finding, unexpected performance outcomes are attributed more to chance or luck than to ability on the part of the ratee (Zuckerman, 1979). Third, observing behavior consistent with what is expected tends to be interpreted as dispositionally caused, whereas unexpected behavior is thought to be more situationally determined.

Two studies that demonstrate the usefulness of attribution theory for understanding performance ratings are, first, Deaux and Emswiller's (1974) study, in which they found that men's successful performance is more likely attributed to their own doing than to chance, while the opposite pattern of attributions is evident for women. The second study, by Scott and Hamner (1975), required raters to evaluate the performance of videotaped actors exhibiting equal mean levels of performance, but with some showing ascending (improving) levels of performance and others descending levels. The actors who showed ascending levels were rated relatively high on motivation and effort and lower on ability as compared to their descending-levels counterparts.

More generally, attribution theory raises the question of what factors raters use in making performance judgments and how those factors influence ratings. For example, when raters attribute poor performance to situational causes, do they give "extra credit," providing higher ratings than warranted on the basis of actual effectiveness, thus allowing for these situational influences? Attribution theory provides some alternative ways of thinking about and studying the performance rating process.

Field Research on Rating Process Issues. Three related process-oriented research approaches address the basic question of what factors "cause" or influence performance ratings and what cues influence raters when they make judgments about others' work performance. These approaches include (a) investigations focused on the effects of rater and ratee characteristics on ratings, (b) exploratory policy capturing research to evaluate the importance of various cues to making summary performance judgments, and (c) confirmatory path analysis studies investigating the impact of selected factors on performance evaluations.

Studies on rater and ratee characteristics reviewed by Landy and Farr (1983), as well as subsequent research, show, first, that rater gender, age, and education have no significant effects on ratings. Second, raters with more experience and knowledge about the job, and also better job performers, provide higher-quality ratings (Mandell, 1956). Third, certain evidence suggests that raters provide somewhat higher ratings for ratees whose race is the same as their own (Kraiger & Ford, 1985; Schmitt & Lappin, 1980), although at least one recent study (Pulakos, White, Oppler, & Borman, 1989) does not confirm this effect. Finally, person perception studies on accuracy in interpreting and predicting the behavior of others (cf. Funder, 1987; Taft, 1955) and studies investigating rater individual difference correlates of performance rating accuracy (Borman, 1979b; Cardy & Kehoe, 1984) suggest that general cognitive ability and field independence, as well as certain temperament constructs such as tolerance, personal adjustment, and self-control, correlate positively with rating accuracy.

Some studies of *ratee* characteristics indicate a sex role stereotype effect for gender, with men being evaluated more highly than women on traditionally male jobs, and a similar advantage for women on female-oriented jobs (Schmitt & Hill, 1977; Schneier & Beusse, 1980). Ratee race findings are as discussed previously, with some evidence of higher evaluations by raters for same-race ratees. Ratee age does not appear to be a significant factor in rating levels. Tenure on the job typically correlates positively with ratings, although not strongly so.

Most of the factors just discussed can affect the *level* of ratings, but another way of examining the influence of these factors is to study differences in *patterns of cues* that raters employ in making ratings. An example would be to evaluate the relative importance attached to factors such as technical competence, interpersonal skill, and position tenure when female workers are being rated compared to ratings of the same factors for their male counterparts. In this example, differences in the patterns of how these factors are used would imply that the judgment processes associated with rating men and women in turn differ.

Two related methods are well suited to investigating this aspect of rating processes. *Policy capturing* (e.g., Christal, 1968; Hobson & Gibson, 1983) and *path analysis* (e.g., James, Mulaik, & Brett, 1982) are appropriate paradigms when scores for ratees are available on both "cue variables," or potentially important factors that might influence ratings, *and* some overall job performance rating for each of these same ratees. In both approaches, scores on the rating factors or independent variables are essentially regressed against the overall performance ratings, and an importance weight for each factor is computed—based on standardized beta weights in the case of policy capturing and on unstandardized regression

weights (structural parameters) or standardized weights (path coefficients) in the case of path analysis.

When policy capturing is employed, analysis is often at the level of the individual rater. A prototype study of this type is that of Hobson, Mendel, and Gibson (1981), who developed 100 profiles of hypothetical professors, each described by scores on 14 dimensions thought to be relevant to performance as a professor (e.g., lecturing delivery and obtaining research funding). Psychology faculty members then rated the overall effectiveness of each hypothetical professor, and regression analyses were conducted for each of these faculty subjects. Conclusions were that different subgroups of faculty subjects had substantially different patterns of regression weights for the dimensions. For example, one subgroup had a reasonably well-balanced set of importance weights across instructional and research aspects of the job, whereas a second subgroup's highest weights were all in the instructional areas, especially lecturing delivery and knowledge of field.

Thus, policy capturing may be useful for identifying the importance that individual raters actually place on different factors when making summary judgments such as overall job performance ratings. One intriguing direction for research with policy capturing would be to group raters initially according to similarities in their patterns of importance weights—that is, presumably, similarities in approaches to integrating this information to make performance judgments. Then these rater "judgment styles" could be related to individual differences, organizational characteristics, or other variables with a theory-driven link so that reasons for similarities and differences in judgment strategies could be explored. Zedeck and Kafry (1977) attempted to examine such a link by grouping nurses according to similarities in patterns of importance weights on performance rating dimensions; then, organization (i.e., hospital) and individual differences (e.g., verbal ability) relationships with rating strategy were examined. No significant relationships were found, but other such attempts to assess links between rater style and individual or organizational variables may shed considerable light on the performance rating process.

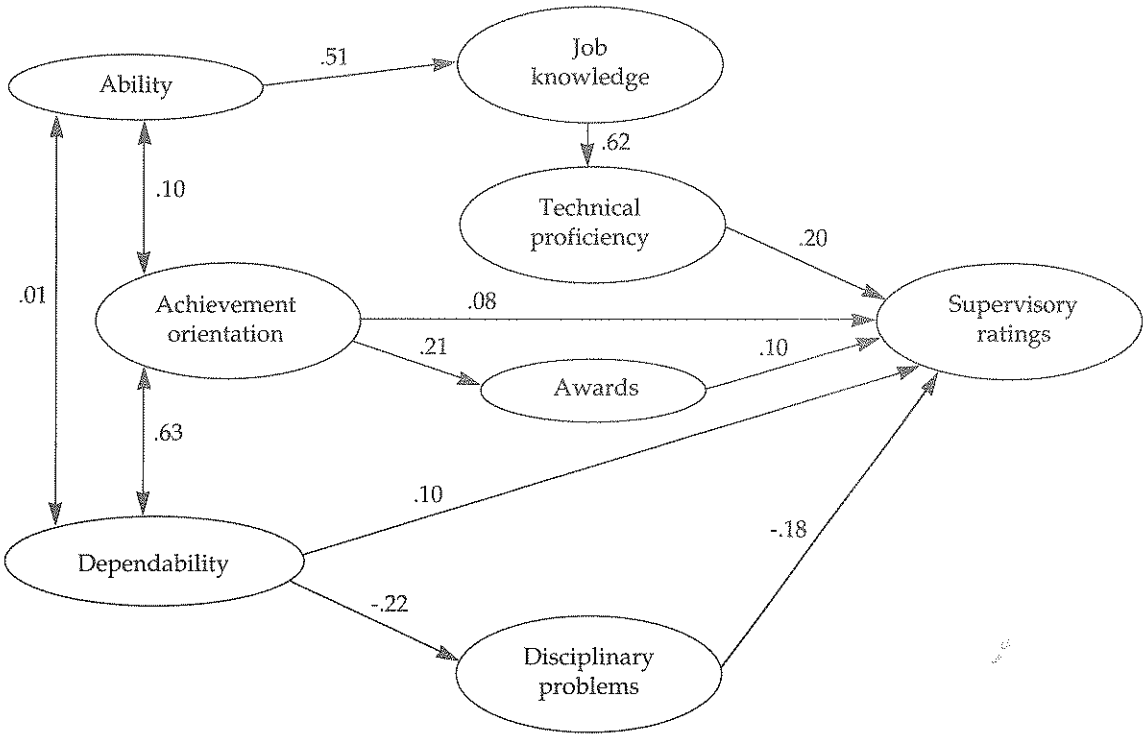
A similar approach to uncovering the factors or cues raters may use to make performance judgments employs path analysis or causal modeling. For example, Hunter (1983) conducted a meta-analysis of 14 studies that used causal analysis to identify relationships between supervisory ratings and (a) general cognitive ability scores, (b) work sample test performance, and (c) job knowledge test scores on the part of ratees. Results suggest that of the three, ratee job knowledge has the largest direct effect on ratings and that cognitive ability has an indirect effect on ratings through its influence on job knowledge.

Responding to Hunter's study, Guion (1983) suggested that other variables besides ratee cognitive ability, job knowledge, and task proficiency might be investigated as having potential effects on performance ratings. Guion saw two sets of variables as especially good candidates for research: interpersonal relationship factors and ratee personal characteristics.

Borman, White, Pulakos, and Oppler (submitted for publication) recently addressed the Guion challenge. Using LISREL VI (Joreskog & Sorbom, 1981), Borman et al. tested the model shown in Figure 4 for a sample of U. S. Army soldiers in nine different jobs. Results of a meta-analysis across the jobs indicated that self-reports of ratee personal characteristics (achievement orientation and dependability) influenced supervisory ratings directly, as well as through their effects on certain performance indicators (number of awards and disciplinary actions received). As in Hunter (1983), general cognitive ability based on *Armed Services Vocational Aptitude Battery* scores for ratees influenced ratings

FIGURE 4

Performance Rating Model Across Nine Army Jobs



Note: $N = 4{,}362$

indirectly through its effect on ratee job knowledge as measured by multiple-choice job knowledge tests. Job knowledge in turn contributed to ratee technical proficiency (work sample performance test scores), with technical proficiency having a direct effect on the supervisors' ratings. Significantly, the variance accounted for in the ratings more than doubled when the variables not included in Hunter's model were added. Thus, we gain more understanding of the determinants of ratings from the extended model here. Models with additional, different variables are needed to more completely map the cues raters use in making performance judgments.

Causal modeling to test hypotheses about relationships between factors potentially influencing performance judgments and the performance ratings themselves seems useful as a strategy for learning more about rating processes. Of course, care should be taken to satisfy measurement requirements in this kind of research. The most important is to avoid including rater perceptions as independent variables in the models (Billings & Wroten, 1978). As an example, strong

relationships between raters' evaluations of ratee characteristics and their performance ratings can be misleading because of common method (the rater) variance between these two measures.

Another contribution of the research into the rating process is that several propositions or hypotheses have emerged from theory-building efforts or surveys of process-related research. Those of most general interest for criterion measurement are:

- Assessment of behavior is a function of the theory and meaning attached by a rater to that behavior and situation. Differences in theory/meaning lead to differences between raters within groups (e.g., supervisors) and between groups (e.g., supervisors vs. self) regarding the interpretation and evaluation of performance, which in turn leads to low interrater agreement.
- Effectiveness of performance appraisal may be enhanced by having observation and recording done by one assessor, and evaluation per se done by an independent assessor; this is in contrast to having one person observe, record, *and* evaluate.
- There is more consistency in the behavior of middle-level performers than there is in the behavior of high- or low-level performers.
- A rater's personal constructs influence the behaviors that are noticed, recorded, and evaluated; behaviors that are not part of or are inconsistent with the rater's personal construct theory are ignored, distorted, or discounted.
- A rater encountering information inconsistent with expectations will be more likely to seek additional information to confirm those expectations.
- Raters will be more likely to recall overall impressions of ratees and the evaluations associated with those impressions than the specific behaviors which gave rise to them.

The first four items in this list appear in Landy, Zedeck, and Cleveland (1983); the fifth and sixth are from De Nisi, Cafferty, and Meglino (1984). These propositions provide a sampling of the kinds of research questions that emerge from the so-called process approach to studying performance ratings. On one hand, they represent basic questions about observation, attention, memory for behavior, integration of performance information, and evaluation. On the other hand, it is hoped that investigating these kinds of questions may also lead to improvements in criterion development and performance measurement procedures.

Enthusiasm for more basic research into rating process issues and hope that the research will result in more error-free and accurate performance measurement is not uniformly shared. Voices of caution and criticism have been heard. Of special concern is the difficulty of generalizing laboratory findings to actual organizational performance rating settings. Ilgen and Favero (1985) identify factors that limit generalizability of most current process-oriented research to *on-line performance appraisal* settings—ratings for purposes of feedback and counseling, for example. However, some of these limitations apply also to our present concerns with criterion development and measurement, typically conducted on a for-research-only basis. Factors relevant here include the following:

- Behavioral observations are not usually made over time on multiple trials in either paper-people or videotape research.
- Situational factors that may reduce (or enhance) performance levels are not well represented in laboratory research.

- Rater-ratee interaction is not properly simulated in paper-people studies or in videotape research.
- Typically, behavior only, not behavior *and* organizational outcomes, is viewed by raters in laboratory rating studies.

According to this analysis, the laboratory context is substantially different from the situation most likely faced by a rater making criterion performance evaluations. More effort is clearly necessary to improve the realism of laboratory research directed toward the study of rating process and related performance rating topics.

It might be noted that the major strategies for examining rating processes do have some similarity in context to real-world personnel evaluation procedures. Subjects working with paper-people protocols are in situations similar to those of supervisors several levels up reviewing behavioral performance reports on employees and making evaluations. Viewing videotape presentations of brief performances places subject raters in conditions similar to those of assessors in an assessment center making effectiveness judgments about performance. The main points remain, however: (a) Caution is in order with respect to generalizing laboratory research on performance ratings, especially to questions of performance appraisal for counseling, feedback, and related administrative purposes, but also to ratings for research only; (b) similar to what has been accomplished for interviewing research (Bernstein, Hakel, & Harlan, 1975), and following Murphy, Herr, Lockhart, & Maguire (1986) in the performance rating area, research should go forward to assess similarities and differences in results between laboratory and field settings; (c) effort and ingenuity should be applied to make laboratory research settings better reflect the organizational settings intended for the research questions asked; and (d) attempts to bring important performance rating research issues out into field settings should continue, with incumbent organization members as subjects.

Objective Criteria

A second major measurement method for criterion performance involves use of objective criteria. Objective criteria employed in personnel research include turnover, absences, production rates, job level and salary, sales, disciplinary cases, and any other directly countable record or index. At first glance, one may presume that objective criteria are better than ratings, which are inherently subjective. Unfortunately, judgment often enters into the assignment of objective criterion scores. Also, objective measures are notoriously deficient as criteria because they typically tap only a small proportion of the job's performance requirements (e.g., Guion, 1965). Contamination can be a problem with some of these criteria as well. Problems such as opportunity bias beyond the assessee's control may influence these outcome measures. Nonetheless, when they are relevant to important conceptual criteria and are reasonably reliable and uncontaminated—or when corrections can be made to reduce contamination—objective measures can be useful in indexing some criterion dimensions.

Turnover. Turnover or attrition is often an important prima facie criterion because the cost of training replacement personnel is usually high; also, having people, especially key people, leave the organization can be disruptive and can adversely affect organizational effectiveness. Turnover is sometimes treated as a single dichotomous variable—a person is either a "leaver" or a "stayer." This treatment fails to distinguish between very different reasons for leaving the organization (e.g., being fired for a disciplinary infraction vs. leaving voluntarily for health reasons). Clearly, turnover for such different reasons will have different patterns of relationships with individual difference or organizational factor

predictors. Prediction of turnover with any substantive interpretation requires a look at the categories of turnover.

Some have advocated two turnover categories, voluntary and involuntary, but in many organizational settings this division is too coarse. Where sample sizes permit, it seems preferable to create a dichotomous variable for each turnover category and then compare on the predictor variable(s) of interest those who left for that reason to all those who stayed. It may be, for example, that employees fired for disciplinary reasons have reliably different scores on certain personality scales—say, lower socialization—compared to stayers, whereas prior health status is the only predictor of leaving the organization for health reasons. This approach to dealing with turnover as a dependent criterion variable, along with research based on turnover models (see Hulin, this volume), appears to offer the most hope for learning more about why individuals leave organizations and what can be done to reduce unwanted turnover.

Absences. For most jobs, having employees at work regularly is important for individual and organizational effectiveness. Accordingly, issues of absences and attendance are legitimate criteria to consider. Unfortunately, three problems plague measures of absences or attendance (Hammer & Landau, 1981). First, criterion contamination is often a factor in measuring absences. Distinctions should be made between involuntary and voluntary absence, for example, with the latter being the more important to predict and then reduce. However, researchers may group involuntary absences for reasons such as legitimate illnesses together with willful, voluntary absences for reasons such as being upset with a supervisor. Making the voluntary-involuntary distinction is often difficult without asking each employee why he or she was absent in each case; even then, one is left with possibly slanted self-reports.

A second problem with absence measures is that they are unstable. In the case of voluntary absences, such factors as the organizational environment can influence absence rates differently for different organization members and can lead to criterion unreliability with absence measures (Hammer & Landau, 1981). One way to reduce this instability is to measure attendance rather than absences. Latham and Pursell (1975) demonstrated that attendance was a more reliable measure than number of absences in a sample of loggers. However, attendance measures necessarily confound voluntary and involuntary absences. A second approach to reducing criterion instability here is to collect absence data over longer time periods (Ilgen & Hollenback, 1977). However, a potential problem is that the antecedent variables or events hypothesized to affect absences may decline in relevance to absences as time goes on, rendering conclusions regarding such hypotheses more and more tenuous over time (Harrison & Hulin, 1989).

Another problem in measuring absences is serious skewing of distributions. Severe truncation, with many sample members having *no* absences, causes difficulties when absences are correlated with scores on another variable such as a predictor measure. Difficulties especially take the form of reduced power for significance tests and depressed correlations with other variables.

A hopeful sign is that emerging alternative strategies for treating absence data, such as event history models, may improve prediction of absences (Harrison & Hulin, 1989) and increase our understanding of the absence-taking process (Fichman, 1989). In sum, voluntary, illegitimate absences should be the focus of study regarding absences in criterion development. This class of absence is presumably a function of employee motivation and willingness to work. Accordingly, it should be predictable from measures of individual differences in employees and organizational factors. Involuntary absences are caused by

events beyond the employee's or organization's control, are typically unpredictable, and are therefore outside the personnel research domain.

Production Rates. For jobs that have observable, countable products that result from individual performance, a production rate criterion is a compelling bottom-line index of performance. However, as often noted (e.g., Guion, 1965; P. C. Smith, 1976), considerable care must be taken in gathering and interpreting production data. For example, work-related dependencies on other employees or on equipment for determining production rates may create bias in these rates. Also, production standards and quota systems (e.g., in data entry jobs) create problems for criterion measurement.

As with absences, instability of production rates is another potential problem. Rothe's (1978) extensive research on production workers doing piecework shows that week-to-week production rates are only moderately reliable. Correlations between successive weeks' production average .75 with incentives, and .53 with no incentives, for increased production (Rothe, 1978). Longer periods for data collection may be necessary to ensure stable criterion production rates. Most importantly, researchers attempting to derive production criteria should pay special attention to possible contaminating influences whereby employees have unequal opportunities to produce at the same rate.

Job Level and Salary. These criteria are intuitively appealing for management and some professional jobs. If in these jobs an organization rewards employees with promotions and salary increases strictly according to their overall performance and worth to the organization, such criteria seem quite appropriate, when adjustments are made for years of service or some similar tenure-related indicator. Regarding adjustments for tenure, Hulin (1962) provided perhaps the most refined treatment, correcting increases in salary for the expected salary rise as predicted by length of service.

Regardless of the corrections made, promotion rate and salary criteria are susceptible to contaminating influences. Situational factors such as timing of higher-level position openings and market value of a particular specialty can adversely affect measures of these criteria. In addition, politics within the organization, when it results in promotion and salary decisions based on factors other than merit, can introduce error into these criterion measures. Finally, as a practical restriction, it is difficult to compare individuals who enter the organization at very different levels and salaries. Promotion rate and salary criteria are best applied in organizations that promote from within.

Nonetheless, provided proper corrections for tenure or experience level can be accomplished and that contaminating factors are not a serious problem, promotions and salary do provide reasonable summary indices of an employee's total worth to the organization. In fact, it might be argued that these criteria reflect a consensus perception (across several supervisors) of an employee's performance in all aspects of the job, weighted according to the organization's value placed on each of these aspects. To increase understanding of work performance per se and to enable evaluation of performance on individual job dimensions, it would of course be preferable to obtain multiple criterion scores for each important dimension of the job.

Sales. Initially, sales jobs may seem ideally suited for the use of objective criteria as performance measures. Total sales volume for a fixed period, number of sales per unit time, or some similar index of bottom-line sales volume appear compelling as global, overall performance measures. Upon closer inspection, however, significant criterion contamination issues are evident for objective sales criteria.

First, summary sales volume measures are a function of both individual skill and effort *and* environmental factors beyond the control of the salesperson. In the context of the J. P. Campbell et al. (1970) behavior-performance-effectiveness model, objective sales volume is an effectiveness measure; where environmental influences are both important and unequal in their effect on salespeople, criterion measurement will be contaminated.

One way to remove contamination is to adjust sales data for factors such as market potential (e.g., Cravens & Woodruff, 1973). A practical strategy for making these adjustments is to create norms for stores, sales territories, or for whatever the appropriate comparison unit is. Then criterion scores for each salesperson can be compared to scores for other salespersons with roughly the same selling-related environment and thus similar opportunities to produce sales.

Unfortunately, an inherent problem with this approach has to do with the norming process itself. For example, if large sales territories with many salespersons are used to accomplish the norming, there may be meaningful differences within territories with respect to opportunity to perform. If smaller territories are used, then the norms tend to be unstable because the mean sales performance comparison indices are based on too few salespersons. Thus, *how* one does the adjusting may be as important as whether or not to adjust. However, the development of norming strategies that overcome these types of problems is likely to be quite useful in criterion development efforts.

Interestingly, sales quotas established for compensation purposes as standards for individual salespersons, or groups of them, can represent an attempt to allow for all environmental contaminants in creating an expected sales performance index. Accordingly, sales volume compared to assigned quota—if the quotas are established with great wisdom about all likely unequal environmental influences that might affect sales but that are beyond the salespersons' control—could actually provide a reasonable global index of sales performance. In practice, quotas are not likely to attain this degree of fairness, and researchers should certainly examine quota development procedures carefully before using them in this manner.

As with most other objective performance measures, sales criteria suffer from problems of deficiency in that global measures of sales volume will often fail to tap important parts of the job. For example, identifying new customers and maintaining good relations with existing customers are important aspects of sales but would not be directly indexed by objective sales measures.

Disciplinary Cases. In the military and in certain highly structured organizations, records of disciplinary actions may provide an index of troublemaking behavior detrimental to the individual's performance and to organizational effectiveness. Care should be taken in interpreting such indices because different supervisors and/or units may have different policies regarding the assignment and recording of disciplinary actions, making comparisons across units difficult. Also, as with absences, the base rate of such actions may be very low and the distributions skewed.

Despite these potential difficulties, some construct validity was obtained for a disciplinary-actions-per-unit-time measure on U.S. Army enlisted personnel early in their careers. The disciplinary actions measure correlated considerably higher with peer and supervisor ratings on dimensions related to personal discipline than it did with ratings of technical skill or physical fitness and military appearance (J. P. Campbell, 1986).

Work Sample Tests

Work sample or performance tests are sometimes developed to provide criteria, especially

for training programs. For example, to help evaluate the effectiveness of training, work samples may be used to assess performance on important tasks before and after training. Such tests can also be used for other personnel research applications, such as criteria in selection studies. Work samples used as criteria should be distinguished from work samples used as predictors of performance in selection. Asher and Sciarrino (1974) and Cascio and Phillips (1979) have reviewed and discussed work samples as selection devices.

Some argue that work sample tests have the highest fidelity for measuring criterion performance. In a sense, the argument is compelling: What could be more direct and fair than to assess employees' performance on a job by having them actually perform some of the most important tasks associated with it? The performance can then be evaluated for level of competence. Yet evaluation of work samples as criteria is not quite so simple, and their use involves several issues—test development issues, conceptual issues regarding their appropriateness, and validity issues. Each is discussed below.

Test Development Issues. The most convincing rationale and procedures for developing work sample tests incorporate the sequence of defining for a job (a) the job content universe, (b) the job content domain, (c) the test content universe, and (d) the test content domain (Guion, 1978). The *job content universe* may consist of an exhaustive list of all tasks and activities carried out by job incumbents in the course of performing the job. A list of these tasks, along with a breakdown of all steps included in each task, provides a good working definition of the job content universe. Next, the *job content domain* is sampled from the content universe. Subject matter experts (SMEs), typically incumbents and/or their supervisors, rate the importance of the tasks in the content universe, and the content domain is identified to contain a workable number of the most critical tasks reflecting all important aspects of the job.

As Guion (1978) points out, the *test content universe* is a theoretical concept intended to include (a) all tasks that might be used in testing, (b) all conditions that might be created for testing individual tasks, and (c) all procedures that could be used to obtain performance scores for testees. From this universe, the *test content domain* is selected, containing not only the tasks to be tested but the context in which each task is to be presented for testing; procedures for generating test scores are also specified. The main issues with testing conditions and context is one of generalizability: Will the testing conditions elicit testee performance that generalizes to actual on-the-job performance? Regarding context, should the task be tested in a work sample or in some other testing mode (e.g., paper-and-pencil job knowledge test or performance ratings)?

As an example of this test development sequence, C. H. Campbell, Campbell, Rumsey, and Edwards (1986) developed a work sample test for evaluating the performance of medical specialists in the U.S. Army. A thorough task-based job analysis first provided a working definition of the job content universe, in which 239 tasks were identified in an exhaustive description of all job content. Also, the tasks were clustered into homogenous groupings on the basis of content. Then SMEs rated the importance of each task for satisfactory performance as a medical specialist, and 30 of the most important tasks, also representative of the task clusters, formed the job content domain.

In addition, the test content universe took into account contextual features that might be built into the testing conditions to make them realistic. Finally, Campbell et al. reviewed components or steps of all surviving tasks to decide on the appropriateness of the performance test mode for tasks. In particular, each task was examined to determine if some hands-on work sample test could be developed to measure performance on the whole task or at least some important components of the task. Fifteen of the 30 tasks could be tested in the hands-on mode, and Campbell et al. designed

performance tests, including appropriate equipment specifications, a scoring system for SMEs to use in evaluating performance on each component or step of each task, and a scorer training program to standardize the assessment of performance on each task. Three example tasks from the performance test for medical specialists are: (a) assemble needle and syringe and draw medication, (b) administer an injection, and (c) initiate an intravenous infusion.

Table 2 then presents six of the eleven task steps for a task in another performance test developed for motor transport operators in the Army (Campbell et al., 1986). A testee's performance on each of these task steps is graded pass or fail based on relatively objective standards, and a percent pass score is used as the total performance test score across all steps.

Another issue in performance test development concerns process versus product: Should the test be focused on the *process* of performing a task or on the *product* that results from completing the task? In general, tasks associated with products (e.g., troubleshooting a problem with a radio) can be oriented toward either product or process; tasks with no resulting products (e.g., interviewing a job candidate) must be scored according to process considerations only. An advantage to scoring products over process is that assessment is typically more objective. However, if the procedures taken to arrive at the product are also important, process assessment is clearly necessary.

Other test development issues relevant to scoring of work samples are germane here. Unscorable or difficult-to-score process steps are to be avoided. For example, checking and inspecting steps are difficult, if not impossible, to observe. Ill-defined steps, such as "adjust protective mask" for the medical specialist example (to what standard?) and complex steps where a testee can do well on one part of the step but poorly on another, should also be avoided. Even for those evaluators who are expert at a job for which a performance test has been developed, training in scoring is critical. The scoring system must be unambiguously understood at an operational level, and evaluators should be instructed to offer the same stimulus set to all testees—that is, to provide the same opportunity for successful performance to each person being tested. Often, for example, evaluators are tempted to coach testees, and this practice should be forbidden.

TABLE 2

Task Steps 3–8 for the Motor Transport Operator Task: Perform Vehicle Emergency and Recovery Procedures

3. Position wooden block on the ground under the axle.
4. Position jack on wooden block.
5. Position jack under the axle housing.
6. Turn out screw jack until jack touches axle housing.
7. Close bleeder valve with jack handle.
8. Raise wheel assembly off the ground.

Still another issue with scoring work samples as part of test development is the relative merits of pass-fail marks versus performance level ratings on test steps. Guion (1978) argues for test step performance ratings because they provide more information. Indeed, many steps seem amenable to a continuous performance scale where such ratings as "more skillful," "faster," and "less waste" may have meaning for evaluating performance. For certain very simple task steps or steps that have definite, straightforward standards, pass-fail may suffice; but it will usually be desirable to develop continuous performance scales for use in work sample testing.

Finally, certain practical concerns with performance test development should be mentioned. Time available for testing will

almost always be limited, so tasks that take a long time to complete may be impractical to test. However, steps from the task may be strategically selected such that they take considerably less time to test as long as the knowledge, skills, and abilities required for the shortened task are highly similar to those required by the entire task. Also, and more obviously, limitations are sometimes posed by unavailability of equipment or impracticality of its use. Very expensive equipment or equipment that might create dangerous conditions (e.g., for a nuclear power plant operator job) make it difficult to develop performance tests for jobs requiring such equipment.

Issues About the Criterion Space Measured by Work Sample/Performance Tests. A second major issue with performance or work sample tests is that researchers may erroneously come to see them as ultimate criteria—that is, these tests are sometimes considered *the* criterion of choice for accurately assessing performance in certain jobs, especially those that require complex motor skills. Performance tests should not be thought of in this light. First, they are clearly maximum performance rather than typical performance measures. As such, they tap the "can-do" more than the "will-do" performance-over-time aspects of effectiveness. Yet "will-do" longer-term performance is certainly important for assessing effectiveness in jobs. Accordingly, these measure are deficient when used exclusively in measuring perforrance.

In summary thus far, inherent shortcomings of work samples for measuring some aspects of performance, as well as practical limitations such as time and equipment constraints, argue against relying on such tests to provide a comprehensive index of overall performance. An important theme of this chapter is that whenever possible, more than one kind of criterion measure should be used, each focusing on that aspect of the criterion it measures best.

Validity Issues. It is beyond the scope of this section to discuss in any depth the different types of validity and their relative merits for evaluating the usefulness of work samples or job knowledge tests. Nonetheless, the view embraced is Dunnette's (Dunnette, 1966): that validation is a process of learning about the meaning of a test's scores and evaluating that meaning.

Content, criterion-related, and construct validity can all take a role in helping to understand what work samples are measuring. For example, content validity notions are relevant to assessing the validity of the tests (Guion, 1978). That is, how closely does a particular test share important content—tasks, components of tasks, conditions—with the content reflected on the job? Actually, content validity can be virtually assured if the work sample test development sequence just outlined is followed conscientiously.

Construct validity principles are more useful in evaluating the meaning of work sample criterion test scores. Examining relationships between these scores and data from other criterion measures can be useful in evaluating the validity of work sample tests. Perhaps the most important construct validation principle applicable in this case is to seek disconfirming evidence for test score validity. For example, strong positive correlations between test scores and ratings of scorer-testee friendship would cast doubt on the validity of work sample scores; likewise, small differences in mean scores between known masters and nonmasters on a target job would certainly reflect poorly on the usefulness of such scores as measures of job proficiency. On the positive side, lack of disconfirming evidence after a series of studies and analyses of this type provides more optimism on the meaning of criterion measures. This, along with evidence of hypothesized positive relationships with other variables, is of course important in building a nomological net in support of construct validity for such measures (Cronbach & Meehl, 1955).

Perhaps an even more compelling framework, seemingly tailor-made for studying the meaning of criterion work sample test scores, is generalizability theory (Cronbach, Rajaratnam, & Gleser, 1963). Generalizability theory can be viewed as an extension of classical psychometric theory. Classical test theory recognizes only a single, undifferentiated source of error; in contrast, generalizability theory recognizes distinct sources of error in measurement. It replaces the reliability coefficient with the coefficient of generalizability, the true score with a more precisely defined universe score, and undifferentiated error variance with specific sources of variance. The universe score is the expected value of the observed score for a person over the universe of items. The coefficient of generalizability is the ratio of universe score variance to observed score variance. In generalizability theory, there may be several different sources of error variance and several coefficients of generalizability, depending on the universe to which the researcher wishes to generalize.

Thus, for any particular measure, it is possible to evaluate the limits of generalizability over different items and conditions of administration. Glaser and Klaus (1962) identified a number of possible elements detrimental to generalizability in performance testing: (a) variations in the testing environment (e.g., with a driving test, variations in weather conditions), (b) instability in the testing equipment (e.g., automatic versus stick shift), and (c) testee attitudes or reactions toward being tested (e.g., varying levels of anxiety concerning the testing). Generalizability designs can be used to evaluate separately each of these possible sources of measurement error in performance testing.

Job Knowledge Tests

Still another major category of criterion measures is the job knowledge test. Job knowledge tests, like work samples, are used primarily as criteria to assess the outcomes of training in organizations. As with work samples, they may also serve as criterion measures for other personnel research purposes. Part of the same content validation sequence, discussed in the last section on work sample/performance tests, can be applied to job knowledge test development (Guion, 1978; Lammlein, 1986). Defining the job content universe with an exhaustive list of nontrivial tasks and identifying the most important tasks using expert judgment is a reasonable way to focus on a workable set of tasks for item writing.

Once the target tasks are identified, items can be prepared, typically in a multiple-choice format, although other kinds of items such as the essay type are of course possible. Just as in writing any other multiple-choice items, care should be taken to ensure that the item stems and response alternatives are clearly stated and that distractor responses are definitely wrong but plausible. Osborn and Campbell (1976) suggest an approach for writing job knowledge items based on important target tasks in the job content domain. For each important task, they suggest asking why the worker fails to perform behaviors correctly on the task. Possibilities are that he or she (a) doesn't know *where* to perform (e.g., where objects, pieces are located), (b) doesn't know *when* to perform a step (problems with sequencing), (c) doesn't know *what* the result should be, or (d) doesn't know *how* to perform individual steps or the entire task. From this framework, items with correct responses, and especially distractor responses, can be productively generated.

An issue with job knowledge test development is when the paper-and-pencil knowledge test medium is appropriate for evaluating job performance. When a task is procedural, requiring primarily knowledge about steps to complete it, and not complex motor skills for performing each step, a job knowledge format seems clearly to be as appropriate as a work sample format. Tasks requiring certain skills

and operations are probably not amenable to job knowledge testing, requiring instead a performance test treatment. Such tasks include (a) those that require finely tuned acts of physical coordination (e.g., a police marksmanship task), (b) those that require quick reaction (e.g., typing a letter under time pressure), and (c) those that require complex time-sharing psychomotor performance (e.g., aircraft cockpit simulator tasks) (Osborn & Campbell, 1976).

As with work sample tests, validity issues for job knowledge tests focus both on the content validity of the test itself and on the validity of inferences made from test scores. Content validity of knowledge tests is aided by a systematic selection of tasks toward which items are then written. The job content universe–content domain–test universe–test domain sequence provides a specific workable procedure for this task selection (Guion, 1978). As discussed, however, care should be taken to represent with job knowledge items only those tasks that can be reasonably tested in a knowledge test mode.

Regarding validity of test scores, one strategy for evaluating their meaning in the spirit of construct validation is to examine empirically alternative hypotheses about relationships between job knowledge test scores and scores on other variables. For example, test items that involve high reading difficulty levels may be biased in favor of persons with better reading skills as opposed to simply measuring job knowledge.

This last example points to a significant potential problem with job knowledge tests targeted toward jobs with low reading level requirements. The reading level of the test should be no higher than that required on the job. Two ideas that have addressed this problem are the "walk-through" testing procedure and an approach to designing picture items for job knowledge tests.

Briefly, the *walk-through test* (Hedge & Teachout, 1986) requires testees to describe what they *would do* in a series of job situations important to the successful conduct of that job. Appropriate equipment is generally available for the testee to "show and tell," thus eliminating the reading requirements present with job knowledge tests. This procedure also allows for the testing of tasks that can't be used in a work sample mode because of problems with safety or inconvenience in actual performance (e.g., certain emergency procedures for a power plant operator).

Regarding the *picture-item test*, Osborn and Ford (1977) developed two interesting versions that replace multiple-choice responses with pictures of equipment. In one version, a picture of a correctly repaired aircraft generator was shown along with pictures of several incorrectly repaired generators, with instructions to select the one that had been correctly repaired. In the second version, testees were asked to identify errors in a picture of an incorrectly completed task. Evaluation studies of the knowledge test component of the walk-through procedures and of knowledge tests using picture items should be carried out.

Developing and Evaluating Models of Job Performance

Considerable theoretical and empirical effort has been dedicated to developing and evaluating individual differences models of predictors. For example, factor analytic studies have led to explication of the aptitude (Guilford, 1967; Thurstone, 1938), personality (Hogan, 1983; Tupes & Christal, 1961), and vocational interest (Rounds & Dawis, 1979) domains. In each of these individual differences areas, the models derived from factor analysis represent attempts to parsimoniously characterize the important dimensions of the domain. With personality, for example, Eysenck's (1947) work suggests that three summary dimensions—introversion/extroversion, neuroticism, and psychoticism—explain much of the variation in temperament. Fiske (1949), Norman (1963),

and more recently Digman and Inouye (1986) report that a five-factor personality dimension system consistently summarizes self-reports and ratings of personality.

With all the effort expended on model development and evaluation on the predictor side, almost no theoretical or empirical work has focused on models of job performance, reflecting the bias of concern in our field for predictors over criteria.

We should now specify what a performance model might look like and how such models could benefit the science and practice of our discipline. A criterion model for a particular job family, for example, might first identify in summary form the important performance requirement constructs for those jobs. The model would also specify relationships between the constructs. In structural modeling terms, the latent structure of the criterion variables would be hypothesized, with one (or preferably more than one) indicator measure identified for each performance construct.

An Example Job Performance Model: The Case of Project A

A characterization of such a model appears in Figure 5. This model, derived from a long-term, large-scale study to improve the selection and classification system for enlisted persons entering the U.S. Army (Project A: J. P. Campbell, 1986; see Hakel, 1986 for a critical review), specifies hypothesized latent criterion variables, along with their respective indicator measures, for nine Army enlisted jobs (J. P. Campbell, McHenry, & Wise, 1990; Wise, Campbell, McHenry, & Hanser, 1986).

The form of this model allows testing it for goodness of fit using confirmatory factor analysis (James, Mulaik, & Brett, 1982; Schmitt & Stults, 1986). It is also possible with the LISREL confirmatory factor analysis program (Joreskog & Sorbom, 1981) to test the generality of a performance model across jobs—in this case, how consistently the hypothesized model was confirmed across the nine jobs. This is the kind of model development effort possible for specifying a summary multidimensional view of the criterion performance requirements for groupings of jobs.

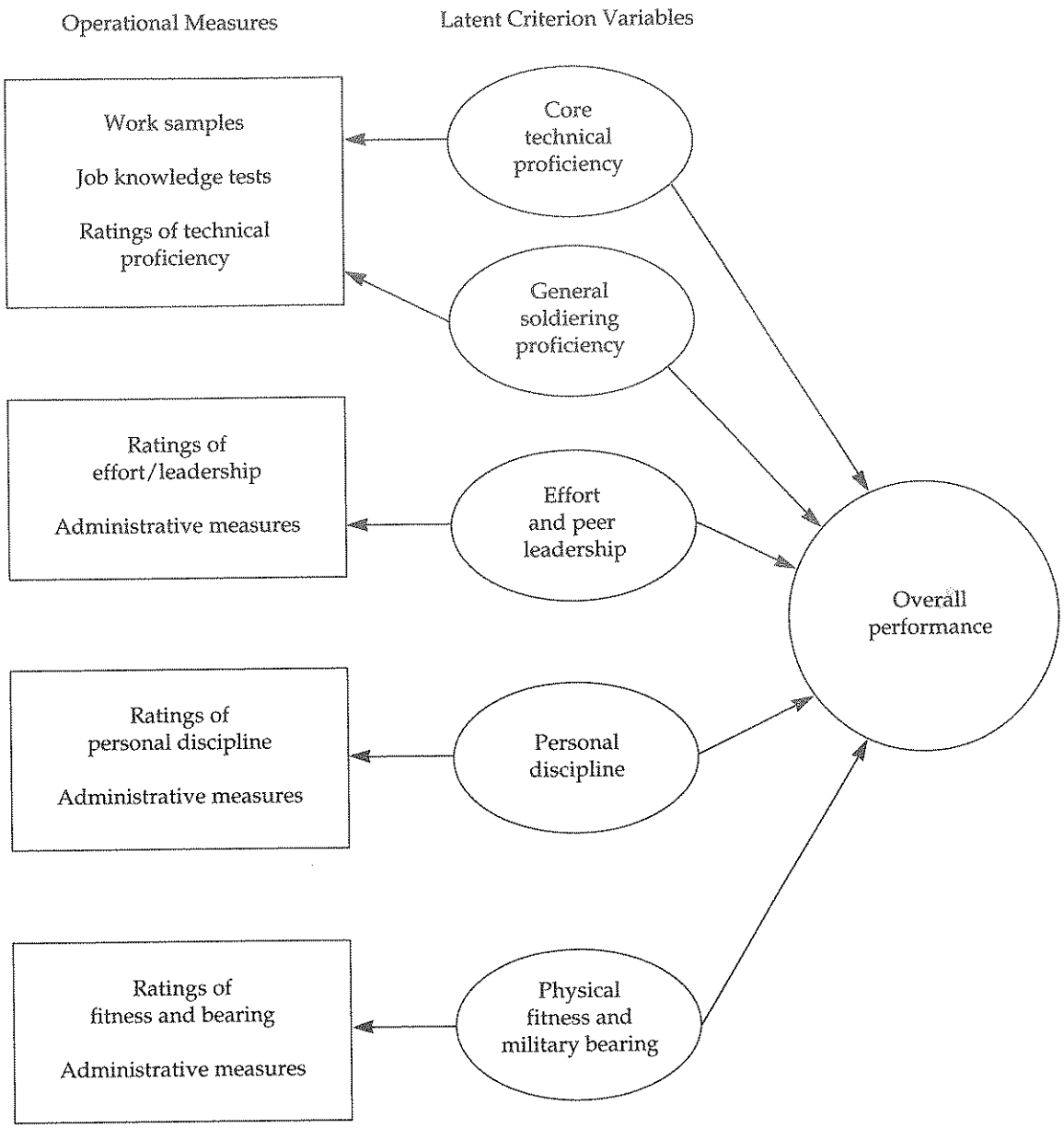
Continuing with the example above, Table 3 shows the main performance indices finally confirmed to measure the five performance constructs. In this example, several of the performance measures were themselves summary indices derived from pilot and field tests of the ratings, work sample, job knowledge, and administrative criterion measures.

In particular, for the 30 job knowledge tasks reflected in the job knowledge test domain for each job (the 15 work sample tasks for each job were a subset of the 30 job knowledge tasks), tasks were grouped according to similarity of task content within each of the nine jobs. The grouping of tasks resulted in from 8 to 15 functional categories for individual jobs (e.g., providing first aid and firing weapons). There was considerable but by no means complete overlap in functional categories across the nine jobs. At this point, work sample test scores and job knowledge test scores were separately generated for each functional category within the jobs by computing a percent passed score for the steps on tasks assigned to that category (work samples) and a percentage-of-items-correct score for steps on the tasks assigned to that job knowledge test category. Then, within each job and type of measure, factor analyses were conducted.

Results showed that some factors consistently emerged across several or even all jobs (e.g., communications/radio operation and vehicle maintenance), and a smaller number of factors were unique to individual jobs. In all, five common work sample and job knowledge test content categories were identified, using a combination of factor analysis results and rational/practical considerations. One technical skills category was retained as unique to each job. At this point, each of the six content

FIGURE 5

Summary Preliminary Model of Enlisted Soldier Performance



From "A Latent Structure Model of Job Performance Factors" by L. Wise et al., 1986. Paper presented at the 94th annual meeting of the American Psychological Association. Copyright 1986 by L. Wise. Reprinted by permission.

TABLE 3

Performance Measures for Each of Five Criterion Constructs

1. Core Technical Proficiency
   - Work samples for specific job
   - Job knowledge test on specific job
2. General Soldiering Proficiency
   - Work samples for parts of job in common with other jobs
   - Job knowledge test on parts of job in common with other jobs
3. Effort and Peer Leadership
   - Peer and supervisory ratings, effort/leadership factor
   - Peer and supervisory ratings, job-specific rating scale factors
   - Administrative awards and certificates measure
4. Personal Discipline
   - Peer and supervisory ratings, personal discipline factor
   - Administrative disciplinary problems measure (–)
   - Administrative promotion rate measure
5. Physical Fitness and Military Bearing
   - Peer and supervisory ratings, fitness/bearing factor
   - Administrative physical readiness measure

categories could be scored for most of the jobs and the work sample and job knowledge test domains were scored separately, employing respectively the percentage of task steps performed correctly for tasks in that category and the percentage of items correct for items in that task category.

For the performance ratings, peer and supervisor ratings on the scales developed to be in common across all jobs (Army-wide scales: e.g., maintaining equipment, following regulations) were first pooled, then intercorrelated and factor analyzed. An interpretable three-factor solution resulted:

- Effort/Leadership—including effort and technical skill in performing the job, peer leadership, and self-development
- Personal discipline—including self-control, integrity, and following regulations
- Military bearing—including physical fitness and military appearance

This three-factor solution reflected very adequately the structure associated with each of the nine jobs. Unit-weighted composites of ratings on the three factors were accordingly computed to represent the Army-wide rating scales in criterion model development efforts. Factor analyses of job-specific rating scales were likewise conducted within each of the nine jobs. Consistently emerging were two factors representing job performance central to the specific content of each job and performance less central to the job content. Unit-weighted composites of pooled peer and supervisor ratings were computed for these two factors as well.

Finally, four administrative measures were judged relevant for performance measurement. Criterion development work suggested that

number of awards and certificates per year in the Army, promotion rate, physical readiness test scores, M-16 rifle qualifying scores, and number of disciplinary actions divided by months in the Army should be target criterion variables. Further, field tests showed that self-reports of these administrative indicators are remarkably accurate. Thus, self-reports of the five indices in Table 3 were employed in the subsequent criterion model development work.

The final set of criterion variables entering into the performance modeling phase for each of the nine jobs appears as follows: [2]

- Two to six work sample content category scores
- Two to six job knowledge content category scores
- Three Army-wide rating category scores
- Two job-specific rating category scores
- One overall effectiveness rating score
- Five administrative measure scores

At this point, correlations between these criterion variables were computed for each job separately. Table 4 displays the intercorrelations for the radio operator job. Exploratory factor analyses were also conducted within each job. One consistent finding across all jobs was that written job knowledge and rating method factors emerged. A second consistent result was that the administrative measures and Army-wide rating factors combined in a conceptually satisfactory way. The awards and certificates variable linked with the effort/leadership rating factor; the disciplinary actions and promotion rate measures clustered together with the Army-wide discipline factor; and the physical readiness test scores and military bearing factor consistently loaded on the same factor. A third finding was that the job-specific rating scale factors loaded more highly on the effort/leadership construct factor than on either of the two proficiency construct factors.

Thus, correlation matrices for each job, follow-up exploratory factor analyses of each of these matrices, and a preliminary model of soldier job performance developed previously (J. P. Campbell & Harris, 1985) led to a target model as shown in Figure 5, except for the inclusion of a written test and a rating factor *and* the adjustment of tying the job-specific rating factors to the effort/leadership factor rather than to the two proficiency factors. This model was tested within each job using the LISREL confirmatory factor analysis procedure.

To summarize the results, chi-square goodness-of-fit tests within each job suggested that the model accounted for the obtained correlations between criterion measures. Further, a LISREL option to test this model across all nine jobs simultaneously showed that the fit was satisfactory. Wise et al. (1986) and Campbell et al. (1990) point out that the model development procedures employed precluded a completely "fair" confirmatory test here. The target model was posited in part on the basis of observed empirical relationships between criteria. Nonetheless, this work remains a very useful illustration of performance model development.

Accordingly, the five-construct, multidimensional representation of job performance appearing in Table 3 seems warranted for each of these jobs. This differentiated depiction of performance was subsequently shown to be important when test validation results demonstrated that cognitive predictors had substantial relationships with the two proficiency factors, and personality predictors correlated more highly with the effort/leadership and discipline factors.

This example shows how a performance model might be derived, tested, and then used in personnel research. How much more widely the model would apply to other jobs is unclear, although it is important to note that the nine jobs studied in this research were selected to be representative of the more

TABLE 4

Correlations Between Criterion Measures for Radio Operators

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Work sample—technical	—																				
2. Work sample—common tasks	25	—																			
3. Work sample—safety	25	18	—																		
4. Work sample—communications	28	27	28	—																	
5. Work sample—vehicle	09	08	16	01	—																
6. Job knowledge—technical	42	31	10	34	11	—															
7. Job knowledge—common tasks	21	31	21	29	09	60	—														
8. Job knowledge—safety	23	18	13	21	10	59	36	—													
9. Job knowledge—communications	21	15	09	38	02	60	50	50	—												
10. Job knowledge—vehicle	22	05	06	21	–06	37	23	28	32	—											
11. Overall performance rating	20	24	15	15	–02	24	17	09	14	03	—										
12. Rating—effort/leadership	24	21	21	15	02	30	28	14	16	06	83	—									
13. Rating—discipline	10	14	07	10	–01	20	15	04	15	06	73	68	—								
14. Rating—military bearing	12	08	10	04	–02	01	02	00	08	–04	64	57	52	—							
15. Rating—core technical	20	24	20	11	–01	29	30	16	15	08	74	81	54	47	—						
16. Rating—other technical	11	18	15	01	00	17	22	08	09	02	66	71	58	40	76	—					
17. Awards/certificates	09	12	06	03	02	10	10	11	–00	–05	17	18	04	11	14	03	—				
18. Physical readiness	01	–10	00	01	–06	–04	–08	04	01	03	11	12	04	34	04	–03	23	—			
19. M-16 qualification	04	05	10	07	05	07	10	08	–04	–06	02	07	–11	–06	05	–03	10	04	—		
20. Disciplinary actions	–09	–03	–07	–12	–03	–16	–09	–13	–20	–10	–31	–31	–32	–25	–16	–17	02	–11	04	—	
21. Promotion rate	08	12	21	09	05	18	17	10	19	13	30	30	26	24	22	22	12	04	03	–34	—

Note: $N = 239$

than 250 entry-level jobs in the U. S. Army. Effort/leadership, personal discipline, and the two technical proficiency constructs appear quite broadly relevant. The military bearing construct seems more specific to a military setting. It would be productive to do more work like this in other job domains to begin to map the performance constructs for a variety of jobs and organizations.

Scientific Implications of Criterion Model Development and Testing

If such models can be confirmed within broad job families, for example, implications for the science of industrial and organizational psychology are considerable. First, the performance requirement constructs of jobs could be more clearly established and delineated by drawing inferences about them from several jobs rather than just one. Second, extensive development and testing of performance models for a variety of jobs might contribute to identification of a way to categorize jobs according to similarities in the performance requirement constructs, leading eventually to a comprehensive mapping of the similarities and differences in these requirements for many kinds of jobs. Third, personnel selection programs could benefit from extended validity generalization efforts in which individual criterion constructs, rather than overall performance, are used within a synthetic validity framework to evaluate the generalizability of test validities. Finally, the reliable and robust representation of performance in multidimensional terms across jobs should considerably improve the understanding of predictor-criterion links by allowing constructive replication of these links in multiple jobs.

Other Observations on Performance Models

Two additional observations should be made concerning the criterion model-building notion. First, criteria for jobs in an organization necessarily reflect the values of that organization. Thus, differences in values between organizations with respect to degree of emphasis to be placed on particular performance areas will somewhat reduce the generalizability of criterion models across organizations. This factor will certainly affect importance weights placed on performance constructs used to form composites for some personnel research applications. Beyond this, an organization's values may influence even the presence or absence of some criterion constructs considered for a performance model in an organization.

Second, method factors are likely to strongly influence the models. For example, ratings tend to correlate more highly with other ratings than they do with work sample performance or objective measure scores, no matter what the criterion construct. This could make the dimensions and structure of a model quite dependent on the criterion measurement methods employed. The only way to address this potential problem is to include the most conceptually appropriate method or methods to measure each criterion construct.

Overall, however, development and evaluation of criterion models is a promising step toward enhanced scientific understanding of criteria. Valid measurement of multidimensional performance on jobs can only improve the science of industrial and organizational psychology. Further, it should work to enhance personnel research practices, where such a multidimensional depiction of performance is desirable.

Conclusions

The following conclusions can be made about criteria and criterion measurement methods.

Performance Ratings

Ratings have the inherent potential advantage of being sensitive to ratee performance over time and across a variety of job situations.

Ideally, raters can average performance levels observed, sampling performance-relevant behavior broadly over multiple trials. Provided observation and subsequent ratings are made on all important dimensions, ratings can potentially avoid problems of contamination and deficiency. A related advantage of ratings is their flexibility for indexing performance on virtually any dimension. If a definition for a dimension can be articulated, then it can form the basis for a rating scale. For example, performance on task proficiency and job knowledge can be rated, as can such dimensions as interpersonal effectiveness, communication, and organization and planning. A third advantage is that ratings avoid artificiality of work samples and job knowledge tests, in that the latter are simulations of the job, whereas ratings are made using actual job performance as input.

Remember that these are *potential* advantages of the rating method. Rater error, bias, and other inaccuracies, well documented here and elsewhere, must be reduced in order for ratings to realize this potential.

Remember, too, that because ratings are used so often in personnel research, it is doubly important that they reach their potential. Both basic and applied research are needed to learn more about what performance ratings are measuring and how to improve on that measurement. Rater training research appears especially promising for reducing rating errors and enhancing the accuracy of evaluations. Research on the performance rating process has not contributed extensively to improvement in rating procedures, but it may yet do so. Field studies of the actual cues that raters use in making performance evaluations should be especially useful.

Objective Measures

Like other methods of measuring criterion performance, objective measures can be quite useful. However, these measures are almost always deficient, contaminated, or both. Indices such as absences, production rates, sales, and disciplinary cases provide data pertinent to only a portion of a job's performance requirements. In addition, some of these indices, along with job level and salary, are often determined in part by factors beyond the employee's control. The latter problem of contamination can sometimes be alleviated by judicious corrections or norming strategies.

More attention should be paid to the psychology of objective measures to increase our understanding of relationships between objective criteria and other measures and to enhance the usefulness of these criteria. One example is to treat reasons for turnover separately (e.g., leaving to take better job vs. leaving for disciplinary reasons), grouping them together only when they reflect the same or similar underlying behavioral meaning. A second example is to form composites with other criterion measures where conceptual rationale and empirical relationships allow. In the Army enlisted personnel performance model (J. P. Campbell et al., 1990), for example, disciplinary actions taken against ratees and ratings of personal discipline were in the same criterion composite; likewise, number of awards and commendations and ratings of effort and leadership appeared in the same criterion cluster.

Work Sample and Job Knowledge Tests

Work sample performance tests should not in any sense be considered as ultimate or even best criteria. Nonetheless, well-conceived and competently constructed performance tests can be valuable measures of maximum performance. If these tests are focused on a good sampling of the important tasks for a job, then they arguably provide a reasonable index of the "can-do" aspect of performance on the job. However, in keeping with Dunnette's (1963) comments criticizing single, overall criterion measures, a single *method* of measuring performance should be discouraged.

Developers of work sample performance tests should follow the Guion (1978) prescription of attending to the job content universe and domain (i.e., sampling tasks carefully) as well as ensuring a realistic set of test conditions, generalizable to on-the-job settings. Analysis of performance test data should, where technically feasible, employ designs from generalizability theory to help pinpoint sources of measurement error (as in test scorers or testing conditions).

Job knowledge tests are typically most appropriate as training criteria, but also for job performance criteria, especially when considerable knowledge of job content and task procedures is a prerequisite for effective on-the-job performance. Of course, job knowledge test scores usually reflect only a modest portion of the total job performance requirement domain.

Final Remarks

Research studies and everyday organizational experience strongly suggest that job performance is multidimensional. Accordingly, criterion development and measurement should proceed on multiple dimensions of job performance. Because jobs are typically defined by clusters of tasks and activities that require similar knowledge, skills, abilities, and personal characteristics, performance on these dimensions will likely be positively correlated (Cooper, 1981) but sufficiently distinct to warrant a multidimensional treatment. If necessary, criteria for a job can be combined to form an appropriately weighted composite after the multiple dimensions have been identified and measured.

The most rational way to conduct criterion development begins with job analysis. Identifying the conceptual criteria in Astin's (1964) nomenclature should come first. An effective strategy here is to use the critical incidents (Flanagan, 1954) or behaviorally anchored rating scale (BARS; Smith & Kendall, 1963) methodology, because these methods naturally produce samplings of actual job behaviors. The content of categories derived from behavioral examples is thus very likely to reflect important conceptual criteria. After these criteria have been identified, the best possible measure for each conceptual criterion should be selected or developed.

Construct validation methods are preferable in evaluating criterion measures (James, 1973; P. C. Smith, 1976). We should attempt to learn as much as possible about what our criterion measures are actually measuring and thus what the criterion scores mean. A good exemplar in this direction is Hunter's (1983) path analysis work, exploring causal links between performance ratings, task proficiency, job knowledge, and general mental ability.

Multitrait-multimethod approaches are also useful for evaluating construct validity; in particular, confirmatory factor analysis procedures show promise (Schmitt & Stults, 1986). Note, however, that not every criterion measurement method should be considered for measuring every performance construct, as in a completely crossed multitrait-multimethod design. Different measurement methods are more and less appropriate for different sets of criteria. For example, job knowledge tests usually don't—and shouldn't—measure the interpersonal aspects of job performance; supervisory ratings may not be usefully employed in evaluating some kinds of specific task proficiency.

Conceptual advances in criterion development over the years have been impressive, and insights into criteria and ideas about how to measure criterion constructs have often been illuminating. The actual measurement of criterion performance is much more problematic. Raters seem limited in their ability to observe behavior and to provide accurate reports of performance. Objective indices are almost invariably deficient, sometimes overly global, and often contaminated, with differential opportunity to score well on them. Work sample performance tests are expensive and

deficient. Job knowledge tests measure only a small proportion of job performance. In some ways, these measurement problems seem insurmountable; for example, the information processing limitations of raters may prove intractable. Yet the effort to improve criterion measurement definitely seems worth it. Criteria provide the foundation for the science of industrial and organizational psychology and the practice of applied personnel research in organizations. Progress is evident, but much more is needed.

Part of this chapter was prepared at the Psychology Department, Ohio State University, while the author was on leave from PDRI. I thank the following Ohio State graduate students for their skilled assistance in reviewing material for the chapter and discussing with me concepts in criterion development during a seminar there: James C. Bassett, Adrienne Colella, Lawrence W. Inks, Laura L. Koppes, Phyllis C. Panzano, and Martha M. Sanders. Thanks also to John P. Campbell, Jeffrey J. McHenry, Lauress L. Wise, and Mike Rumsey for reading and commenting on sections of the chapter, to Marvin D. Dunnette and Patricia C. Smith for providing feedback on an earlier draft, and finally to Kim Downing and John Novak for organizing and typing several versions of this manuscript.

Notes

1 Accuracy of ratings can be distinguished from their validity. *Accuracy* is typically thought to require both the correct rank order of ratees *and* the correct absolute level against some set of target scores. *Validity* is concerned with the correct rank order only. This convention will be followed, although, as we will see, a number of different conceptualizations and operational definitions of accuracy somewhat complicate this depiction of accuracy.

2 There are two comments about this listing. First, to simplify description of the criterion model development analyses, two sets of measures considered in the J. P. Campbell et al. (1990) analyses are not included in the array: (a) experimental combat performance prediction ratings and (b) training knowledge test measures. Second, the two to six work sample or job knowledge category scores mean that some jobs did not have tasks relevant to some work sample/job knowledge categories. For example, administrative specialists have no communications/radio operation and vehicle maintenance tasks and therefore have missing data for those work sample and job knowledge categories.

References

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist, 36,* 715–729.

Ackerman, P. L. (1987). Individual differences in skilled learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102,* 3–27.

Adams-Webber, J. R. (1979). *Personal construct theory.* New York: Wiley.

Althauser, R. P. (1974). Inferring validity from the multitrait-multimethod matrix: Another assessment. In H. L. Costner (Ed.), *Sociological methodology 1973–1974* (pp. 106–107). San Francisco: Jossey-Bass.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27,* 519–533.

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

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology, 72,* 567–572.

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

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

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

Bartlett, C. J. (1983). What's the difference between valid and invalid halo? Forced choice measure-