4

CRITERIA

Definitions, Measures, and Evaluation

LEARNING GOALS

By the end of this chapter, you will be able to do the following:

- 4.1 Define and identify criteria—yardsticks used to assess employee success
- 4.2 Distinguish the static, dynamic, and individual dimensions of criteria and their implications
- 4.3 Define and measure contextual, task, and counterproductive behaviors
- 4.4 Develop criteria that address challenges such as job performance unreliability, unreliability in the observation of performance, and the multidimensionality of performance
- 4.5 Consider the importance of situational determinants of performance and develop and evaluate criteria using standards such as relevance, sensitivity, and practicality
- 4.6 Develop criteria that will minimize the detrimental impact of criterion deficiency and contamination and choose whether to use composite or multiple criteria
- 4.7 Distinguish observed from unobserved criteria and their antecedents
- 4.8 Consider nonnormal distributions of performance and their implications in terms of the presence and production of star performers

The development of criteria that are adequate and appropriate is at once a stumbling block and a challenge to the HR specialist. Behavioral scientists have bemoaned the "criterion problem" through the years. The term refers to the difficulties involved in the process of conceptualizing and measuring performance constructs that are multidimensional, dynamic, and appropriate for different purposes (Austin & Villanova, 1992). Yet the effectiveness and future progress of knowledge with respect to most HR policies and interventions depend fundamentally on our ability to resolve this baffling question.

The challenge is to develop theories, concepts, and measurements that will achieve the twin objectives of enhancing the utility of available procedures and programs and deepening our understanding of the psychological and behavioral processes involved in job performance. Ultimately, our goal is to develop a comprehensive theory of the performance of men and women at work (Campbell & Wiernik, 2015; Viswesvaran & Ones, 2000).

In the early days of applied psychology, according to Jenkins (1946), most researchers and practitioners tended to accept the tacit assumption that criteria were either given by God or just to be found lying about. It is regrettable that even today we often resort to the most readily available or most expedient criteria when, with a little more effort and thought, we could probably develop much better ones. Nevertheless, progress has been made as the field has come to recognize that criterion measures are samples of a larger performance universe and that as much effort should be devoted to understanding and validating criteria as is devoted to identifying predictors of them (Campbell & Wiernik, 2015). Wallace (1965) expressed the matter aptly when he said that the answer to the question "Criteria for what?" must certainly include "for understanding" (p. 417). Let's begin by defining our terms.

DEFINITION

Criteria has been defined from more than one point of view. From one perspective, criteria are standards that can be used as yardsticks for measuring employees' degree of success on the job (Bass & Barrett, 1981; Guion, 1965; Landy & Conte, 2016). This definition is quite adequate within the context of personnel selection, placement, promotion, succession planning, and performance management. It is useful when prediction is involved—that is, in the establishment of a functional relationship between one variable, the predictor, and another variable, the criterion. However, there are times when we simply wish to evaluate without necessarily predicting. Suppose, for example, that the HR department is concerned with evaluating the effectiveness of a recruitment campaign aimed at attracting members of underrepresented groups (e.g., women for science, technology, engineering, and mathematics—STEM—positions). Various criteria must be used to evaluate the program adequately. The goal in this case is not prediction but rather evaluation. Fundamentally, one distinction between predictors and criteria is time (Mullins & Ratliff, 1979). For example, if evaluative standards such as written or performance tests are administered *before* an employment decision is made (i.e., to hire or to promote), the standards are labeled predictors. If evaluative standards are administered *after* an employment decision has been made (i.e., to evaluate performance effectiveness), the standards are labeled criteria.

This discussion leads to the conclusion that a more comprehensive definition is required, regardless of whether we are predicting or evaluating. As such, a more general definition is that a criterion represents something important or desirable. It is an operational statement of the goals or desired outcomes of the program under study (Astin, 1964). It is an **evaluative standard** that can be used to measure a person's performance, attitude, motivation, and so forth (Blum & Naylor, 1968). Examples of criteria are presented in Table 4.1, which has been modified from those offered by Dunnette and Kirchner (1965), Guion (1965), and others (e.g., Aguinis, O'Boyle, Gonzalez-Mulé, & Joo, 2016; Brock, Martin, & Buckley, 2013). Although many of these measures often would fall short as adequate criteria, each of them deserves careful study in order to develop a comprehensive sampling of job or program performance. There are several other requirements of criteria in addition to desirability and importance, but before examining them, we must first consider the use of job performance as a criterion.

TABLE 4.1 ■ Possible Measures of Criteria

Output Measures
Commission earnings
Dollar volume of sales
Number of candidates attracted (recruitment program)
Number of items sold
Number of letters typed
Number of new patents (or creative/innovative inventions and projects)
Number of publications in scientific journals
Readership of an advertisement
Units produced
Quality Measures
Cost of spoiled or rejected work
Number of complaints and dissatisfied persons (clients, customers, subordinates, colleagues)
Number of errors (coding, filing, bookkeeping, typing, diagnosing)
Number of errors detected (inspector, troubleshooter, service person)
Number of policy renewals (insurance sales)
Rate of scrap, reworks, or breakage
Lost Time
Employee turnover (individual-, team-, and unit-level turnover)
Frequency of cyberloafing
Frequency of non-work-related e-mail sent and received at work
Frequency of long coffee or smoke breaks taken without approval
Length and frequency of unauthorized pauses
Length of service
Number of discharges for cause
Number of occasions (or days) absent
Number of times tardy
Number of transfers due to unsatisfactory performance
Number of voluntary quits
Employability, Trainability, and Promotability
Length of time between promotions
Level of proficiency reached in a given time
Number of promotions in a specified time period
Number of times considered for promotion
Rate of salary increase
Time to reach standard performance

Ratings of Performance
Ratings of behavioral expectations
Ratings of performance in simulations and role-playing exercises
Ratings of performance in work samples
Ratings of personal traits or characteristics
Ratings of skills
Counterproductive Behaviors
Abuse toward others (e.g., bullying)
Disciplinary transgressions
Military desertion
Personal aggression
Political deviance
Property damage
Sabotage
Substance abuse
Theft

JOB PERFORMANCE AS A CRITERION

Based on the measures of criteria in Table 4.1, we see that *performance* can be defined as what people do or what people produce. Interestingly, although performance is one of the most central constructs in applied psychology, it may be defined in terms of behavior or results (DeNisi & Smith, 2014). For example, Campbell and Wiernik (2015), Aguinis (2019), and Beck, Beatty, and Sackett (2014) *defined performance based on employee behaviors and actions*—particularly those that are relevant to organizational goals. In fact, Campbell and Wiernik (2015) were quite forceful in their conclusion that "performance should be specified in behavioral terms as things that people do" (p. 67). By contrast, Minbashian and Luppino (2014), O'Boyle and Aguinis (2012), and Aguinis et al. (2016) *defined performance in terms of results—what people produce.* Given the coexistence of these definitions, it is not surprising that some researchers such as Viswesvaran and Ones (2000) defined *performance* as *both* behavior and results, as follows: "scalable actions, behavior and outcomes that employees engage in or bring about that are linked with and contribute to organizational goals" (p. 216).

Although there are different proponents for the behavior- and results-based definitions, the two are clearly related. For example, behaviors such as exerting more effort at work (i.e., behavior-based performance) are likely to result in more and better outcomes (i.e., results-based performance). In fact, the empirical evidence shows that these two types of performance are distinct but also related at nontrivial levels (e.g., Beal, Cohen, Burke, & McLendon, 2003; Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995). So, the question is not whether to define *performance* as behaviors or results, but when and why to define *performance* one way or another (or both).

Some of the proponents of the behavior-based definition of *performance* believe that there is, nevertheless, value in defining it as results. For example, although Beck et al. (2014) adopted the behavior-based approach, they clarified that a results-based definition "may indeed serve many useful organizational and research purposes" (p. 534). Specifically, Aguinis and O'Boyle (2014) chose the results-based definition of *performance* because "a focus on results rather than behaviors is most appropriate when (a) workers are skilled in the needed behaviors, (b) behaviors and results are obviously related, and (c) there are many ways to do the job right" (p. 316). These researchers and others have adopted a results-based definition also because it plays a central role regarding organizational-level outcomes (Boudreau & Jesuthasan, 2011; Cascio & Boudreau, 2011a). In other words, in terms of assessing firm performance, we are more interested in *what* employees produce than in *how* they produce these results.

The term **ultimate criterion** (Thorndike, 1949) describes the full domain of performance and includes everything—all behaviors and results—that ultimately define success on the job. Such a criterion is ultimate in the sense that one cannot look beyond it for any further standard by which to judge performance. The ultimate criterion of a salesperson's performance must include, for example, the quality of customer interactions; time spent with customers; knowledge of products; total sales volume; total number of new accounts brought in during a particular time period; amount of customer loyalty built up by the salesperson; total amount of his or her influence on the morale or sales records of other company salespersons; and overall effectiveness in planning activities and calls, controlling expenses, and handling necessary reports and records. In short, the ultimate criterion is a concept that is strictly conceptual and, therefore, cannot be measured or observed; it embodies the notion of "true," "total," "long-term," or "ultimate worth" to the employing organization.

Although the ultimate criterion is stated in broad terms that often are not susceptible to quantitative evaluation, it is an important construct because the relevance of any operational criterion measure and the factors underlying its selection are better understood if the conceptual stage is clearly and thoroughly documented (Astin, 1964).

DIMENSIONALITY OF CRITERIA

Operational measures of the conceptual criterion may vary along several dimensions. In a classic article, Ghiselli (1956) identified three different types of criterion dimensionality: static, dynamic, and individual dimensionality. We examine each of these three types of dimensionality next.

Static Dimensionality

If we observe job performance at any single point in time, we find that it is multidimensional in nature. This type of multidimensionality refers to two issues: (1) Individuals may be high on one performance facet and simultaneously low on another, and (2) a distinction is needed between maximum and typical performance.

Regarding the various performance facets, Rush (1953) found that a number of relatively independent skills are involved in selling. Thus, a salesperson's learning aptitude (as measured by sales school grades and technical knowledge) is unrelated to objective measures of his or her achievement (such as average monthly volume of sales or percentage of quota achieved), which, in turn, is independent of the salesperson's general reputation (e.g., planning of work, rated potential value to the firm), which, in turn, is independent of his or her sales techniques (e.g., sales approaches, interest and enthusiasm).

In broader terms, we can consider two general facets of performance: **task performance** and **contextual performance** (Borman & Motowidlo, 1997). Contextual performance has also been labeled **pro-social behaviors** or **organizational citizenship performance** (Borman, Brantley, & Hanson, 2014). Task performance and contextual performance do not necessarily go hand in hand (Bergman, Donovan, Drasgow, Overton, & Henning, 2008). An employee can be highly proficient at her task, but be an underperformer with regard to contextual performance (Bergeron, 2007). Task performance is defined as (a) activities that transform raw materials into the goods and services that are produced by the organization and (b) activities that help with the transformation process by replenishing the supply of raw materials; distributing its finished products; or providing important planning, coordination, supervising, or staff functions that enable it to function effectively and efficiently (Cascio & Aguinis, 2001). Contextual performance is defined as those behaviors that contribute to the organization's effectiveness by providing a good environment in which task performance can occur. Contextual performance includes behaviors such as the following:

- Persisting with enthusiasm and exerting extra effort as necessary to complete one's own task activities successfully (e.g., being punctual and rarely absent, expending extra effort on the job)
- Volunteering to carry out task activities that are not formally part of the job (e.g., suggesting organizational improvements, making constructive suggestions)
- Helping and cooperating with others (e.g., assisting and helping coworkers and customers)
- Following organizational rules and procedures (e.g., following orders and regulations, respecting authority, complying with organizational values and policies)
- Endorsing, supporting, and defending organizational objectives (e.g., exhibiting organizational loyalty, representing the organization favorably to outsiders)

Researchers have more recently identified what some consider to be the "dark side" of contextual performance, often labeled **workplace deviance** or **counterproductive behaviors** (Marcus, Taylor, Hastings, Sturm, & Weigelt, 2016; Spector et al., 2006). Although contextual performance and workplace deviance are seemingly at the opposite ends of the same continuum, evidence suggests that they are distinct from each other (Judge, LePine, & Rich, 2006; Kelloway, Loughlin, Barling, & Nault, 2002). In general, workplace deviance is defined as voluntary behavior that violates organizational norms and thus threatens the well-being of the organization, its members, or both (Robinson & Bennett, 1995). Vardi and Weitz (2004) identified over 100 such "organizational misbehaviors" (e.g., alcohol/drug abuse, belittling opinions, breach of confidentiality), and several scales are available to measure workplace deviance based on self- and other reports (Bennett & Robinson, 2000; Blau & Andersson, 2005; Hakstian, Farrell, & Tweed, 2002; Kelloway et al., 2002; Marcus, Schuler, Quell, & Hümpfner, 2002; Spector et al., 2006; Stewart, Bing, Davison, Woehr, & McIntyre, 2009). Some of the self-reported deviant behaviors measured by these scales are the following:

- Exaggerating hours worked
- Falsifying a receipt to get reimbursed for more money than was spent on business expenses

- Starting negative rumors about the company
- Gossiping about coworkers
- Covering up one's mistakes
- Competing with coworkers in an unproductive way
- Gossiping about one's supervisor
- Staying out of sight to avoid work
- Taking company equipment or merchandise
- Blaming one's coworkers for one's mistakes
- Intentionally working slowly or carelessly
- Being intoxicated during working hours
- Seeking revenge on coworkers
- Presenting colleagues' ideas as if they were one's own

Regarding the typical versus maximum performance distinction, **typical performance** refers to the average level of an employee's performance, whereas **maximum performance** refers to the peak level of performance an employee can achieve (DuBois, Sackett, Zedeck, & Fogli, 1993; Sackett, Zedeck, & Fogli, 1988). Employees are more likely to perform at maximum levels when they understand they are being evaluated, when they accept instructions to maximize performance on the task, and when the task is of short duration. A meta-analysis that included 42 studies and a total sample size of 4,129 workers found that the average observed correlation between measures of maximum performance (i.e., what employees *can* do) with measures of typical performance (i.e., what employees *will* do) is only .33 (Beus & Whitman, 2012). The distinction between maximum and typical performance is a fairly new development, so research regarding this topic is still nascent. Nevertheless, based on the few studies available, the meta-analytic evidence shows that general mental ability is more strongly correlated to maximum performance (r = .25) than typical performance (r = .16).

Unfortunately, research and HR practices on criteria frequently ignore the fact that job performance often includes many facets that are relatively independent, such as task and contextual performance and the important distinction between typical and maximum performance. Because of this, employee performance is often not captured and described adequately. In addition, to capture the performance domain in a more exhaustive manner, researchers should also pay attention to the temporal dimensionality of criteria.

Dynamic or Temporal Dimensionality

Once we have defined clearly our conceptual criterion, we must then specify and refine operational measures of criterion performance (i.e., the measures actually to be used). Regardless of the operational form of the criterion measure, it must be taken at some point in time. When is the best time for criterion measurement? Optimum times vary greatly from situation to situation, and conclusions therefore need to be couched in terms of when criterion measurements were taken. Far different results may occur depending on when criterion measurements were taken (Weitz, 1961), and failure to consider the temporal dimension may lead to misinterpretations.

In predicting the short- and long-term success and survival of life insurance agents, for example, ability as measured by standardized tests is significant in determining early sales

success, but interests and personality factors play a more important role later on (Ferguson, 1960). The same is true for accountants (Bass & Barrett, 1981). Thus, after two years as a staff accountant with one of the major accounting firms, interpersonal skills with colleagues and clients are more important than pure technical expertise for continued success. In short, criterion measurements are not independent of time.

Temporal dimensionality is a broad concept because criteria may be "dynamic" in three distinct ways: (1) changes over time in average levels of group performance, (2) changes over time in validity coefficients, and (3) changes over time in the rank ordering of scores on the criterion (Barrett, Caldwell, & Alexander, 1985).

Regarding *changes in group performance over time,* Ghiselli and Haire (1960) followed the progress of a group of investment salespeople for 10 years. During this period, they found a 650% improvement in average productivity, and still there was no evidence of leveling off! However, this increase was based only on those salespeople who survived on the job for the full 10 years; it was not true of *all* salespeople in the original sample. To be able to compare the productivity of the salespeople, their experience must be the same, or else it must be equalized in some manner (Ghiselli & Brown, 1955). Indeed, a considerable amount of other research evidence cited by Barrett, Caldwell, and Alexander (1985) does not indicate that average productivity improves significantly over lengthy time spans.

Criteria also might be dynamic if the *relationship between predictor* (e.g., preemployment test scores) *and criterion scores* (e.g., supervisory ratings) *fluctuates over time* (e.g., Jansen & Vinkenburg, 2006). Bass (1962) found this to be the case in a 42-month investigation of salespeople's rated performance. He collected scores on three ability tests, as well as peer ratings on three dimensions, for a sample of 99 salespeople. Semiannual supervisory merit ratings served as criteria. The results showed patterns of validity coefficients for both the tests and the peer ratings that *appeared* to fluctuate erratically over time. However, he reached a much different conclusion when he tested the validity coefficients statistically. He found no significant differences for the validities of the ability tests, and when peer ratings were used as predictors, only 16 out of 84 pairs of validity coefficients (roughly 20%) showed a statistically significant difference (Barrett et al., 1985).

Researchers have suggested two hypotheses to explain why *validities might change over time.* One, the **changing task model,** suggests that although the relative amounts of ability possessed by individuals remain stable over time, criteria for effective performance might change in importance. Hence, the validity of predictors of performance also might change. The second model, known as the **changing subjects model,** suggests that although specific abilities required for effective performance remain constant over time, each individual's level of skills and ability changes over time, and that is why validities might fluctuate (Henry & Hulin, 1987). Neither model has received unqualified support.

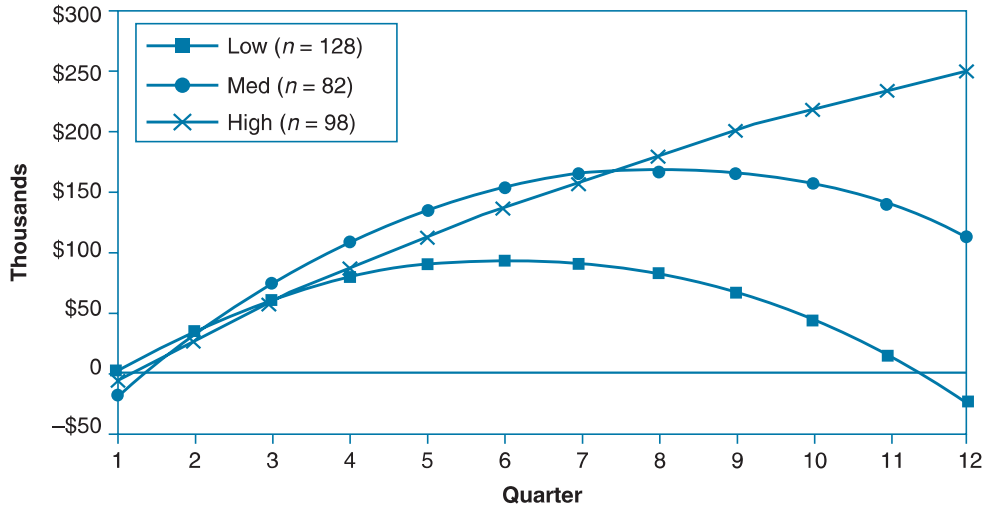
The third type of criteria dynamism addresses possible *changes in the rank ordering of scores over time.* This form of dynamic criteria has attracted substantial attention (e.g., Hofmann, Jacobs, & Baratta, 1993; Hulin, Henry, & Noon, 1990) because of the implications for the conduct of validation studies and personnel selection in general. If the rank ordering of individuals on a criterion changes over time, future performance becomes a moving target. Under those circumstances, it becomes progressively more difficult to predict performance accurately the farther out in time from the original assessment. Do performance levels show systematic fluctuations across individuals? The answer seems to be yes because the preponderance of evidence suggests that prediction deteriorates over time (Keil & Cortina, 2001). Overall, correlations among performance measures collected over time show what is called a "simplex" pattern of higher correlations among adjacent pairs and lower correlations among measures taken at greater time intervals (e.g., the correlation between month 1 and month 2 is greater than the correlation between month 1 and month 5) (Steele-Johnson, Osburn, & Pieper, 2000).

Deadrick and Madigan (1990) collected weekly performance data from three samples of sewing machine operators (i.e., a routine job in a stable work environment). Results showed the simplex pattern such that correlations between performance measures over time were smaller when the time lags increased. Deadrick and Madigan concluded that relative performance is not stable over time. A similar conclusion was reached by Hulin et al. (1990), Hofmann et al. (1993), and Keil and Cortina (2001): Individuals seem to change their rank order of performance over time (see Figure 4.1). In other words, there are meaningful differences in intraindividual patterns of changes in performance across individuals, and these differences are also likely to be reflected in how individuals evaluate the performance of others (Reb & Cropanzano, 2007).

The recent development of wearable sensors and other technological advancements allows applied researchers to capture individuals' fluctuations in performance over time more precisely (Chaffin et al., 2017; Tomczak, Lanzo, & Aguinis, 2018). **Wearable sensors** are mobile devices containing electronic components that are able to gather real-time data on the device-bearing person and his or her context (similar to devices such as Fitbit and Jawbone). For example, individuals can carry a smartphone fitted with a microphone and Bluetooth modules that can generate data including ambient sound and proximity to other devices and interactions with other people (e.g., customers, coworkers). A wearable sensor can also record an employee's location via GPS. Also, individuals can receive a text message asking them to answer questions about where they are and what they are doing using their smartphones—and that information can be correlated with physiological markers (e.g., blood pressure, heart rate) and affective and attitudinal variables (e.g., satisfaction, emotions) (Beal, 2015).

Taken together, these technologies allow organizations to implement **employee monitoring systems** that capture employee performance on an ongoing basis and can capture performance fluctuations, and possible reasons and outcomes of those fluctuations, on a monthly, weekly, daily, and even hourly basis. Although still in their infancy, these technological advances are opening up entire new research avenues regarding what is labeled

FIGURE 4.1 ■ Regression Lines for Three Ordinary Least Squares Clusters of Insurance Agents—Low, Moderate, and High Performers—Over Three Years



Source: Hoffman, D. A., Jacobs, R., & Baratta, J. E. (1993). Dynamic criteria and the measurement of change. *Journal of Applied Psychology, 78,* 194–204.

intraindividual performance fluctuations, or a **within-person performance analysis.** Wearable sensors allow for the collection of **big data** (i.e., as has been done in other fields such as computer science and genomics) that was unthinkable just a few years ago (Harlow & Oswald, 2016). These new types of measures can help us understand whether fluctuations, for example, in daily organizational citizenship behaviors, are related to factors related to the work per se or to the job's social environment (e.g., Spence, Ferris, Brown, & Heller, 2012).

Overall, a major conclusion from recent research efforts is that within-person variability in performance is not necessarily the result of faulty measures (Dalal, Bhave, & Fiset, 2014). Accordingly, an important question posed by these findings is, How can we possibly predict performance if it is a moving target? The answer is that, in predicting performance, it is necessary to take the time dimension into account. Specifically, the goal then is to predict performance within a prespecified time span. For example, there is a need to understand which types of measures predict short- versus mid- versus long-term performance.

Individual Dimensionality

It is possible that individuals performing the same job may be considered equally good, yet the nature of their contributions to the organization may be quite different. Thus, different criterion dimensions should be used to evaluate them. Kingsbury (1933) recognized this problem almost 90 years ago when he wrote:

> Some executives are successful because they are good planners, although not successful directors. Others are splendid at coordinating and directing, but their plans and programs are defective. Few executives are equally competent in both directions. Failure to recognize and provide, in both testing and rating, for this obvious distinction is, I believe, one major reason for the unsatisfactory results of most attempts to study, rate, and test executives. Good tests of one kind of executive ability are not good tests of the other kind. (p. 123)

Although in the managerial context described by Kingsbury there is only one job, it might plausibly be argued that in reality there are two (i.e., directing and planning). The two jobs are qualitatively different only in a psychological sense. In fact, the study of individual criterion dimensionality is a useful means of determining whether the same job, as performed by different people, is psychologically the same or different.

CHALLENGES IN CRITERION DEVELOPMENT

Competent criterion research is one of the most pressing needs of personnel psychology today—as it has been in the past. Stuit and Wilson (1946) demonstrated that continuing attention to the development of better performance measures results in better predictions of performance. The validity of these results has not been dulled by time (Viswesvaran & Ones, 2000). In this section, therefore, we consider three types of challenges faced in the development of criteria, point out potential pitfalls in criterion research, and sketch a logical scheme for criterion development.

At the outset, it is important to set certain "chronological priorities." First, criteria must be developed and analyzed, for only then can predictors be constructed or selected to predict relevant criteria. Far too often, unfortunately, predictors are selected carefully, followed by a hasty search for "predictable criteria." To be sure, if we switch criteria, the validities of the predictors will change, but the reverse is hardly true. Pushing the argument to its logical extreme,

if we use predictors with no criteria, we will never know whether or not we are selecting those individuals who are most likely to succeed. Observe the chronological priorities! At least in this process we know that the chicken comes first and then the egg follows.

In the sections that follow, we address four basic challenges in criterion development (Ronan & Prien, 1966, 1971): reliability of performance, reliability of performance observation, dimensionality of performance, and modification of performance by situational characteristics. Let's consider the first three in turn. The fourth is the focus of the section "Performance and Situational Characteristics."

Challenge #1: Job Performance (Un)Reliability

Job performance reliability is a fundamental consideration in HR research and practice, and its assumption is implicit in all predictive studies. *Reliability* in this context refers to the consistency or stability of job performance over time. Are the best (or worst) performers at time 1 also the best (or worst) performers at time 2? As noted in the previous section, the rank order of individuals based on job performance scores does not necessarily remain constant over time.

Thorndike (1949) identified two types of unreliability—intrinsic and extrinsic—that may shed some light on the issue. **Intrinsic unreliability** is due to personal inconsistency in performance, whereas **extrinsic unreliability** is due to sources of variability that are external to job demands or individual behavior. Examples of the latter include variations in weather conditions (e.g., for outside construction work); unreliability due to machine downtime; and, in the case of interdependent tasks, delays in supplies, assemblies, or information. Much extrinsic unreliability is due to careless observation or poor control.

Faced with all of these potential confounding factors, what can be done? One solution is to **aggregate** (average) behavior over situations or occasions, thereby canceling out the effects of incidental, uncontrollable factors. To illustrate this, Epstein (1979, 1980) conducted four studies, each of which sampled behavior on repeated occasions over a period of weeks. Data in the four studies consisted of self-ratings, ratings by others, objectively measured behaviors, responses to personality inventories, and psychophysiological measures such as heart rate. The results provided unequivocal support for the hypothesis that stability can be demonstrated over a wide range of variables so long as the behavior in question is averaged over a sufficient number of occurrences. Once adequate performance reliability was obtained, evidence for validity emerged in the form of statistically significant relationships among variables. Similarly, Martocchio, Harrison, and Berkson (2000) found that increasing aggregation time enhanced the size of the validity coefficient between the predictor, employee lower-back pain, and the criterion, absenteeism.

Two further points bear emphasis. One, there is no shortcut for aggregating over occasions or people. In both cases, it is necessary to sample adequately the domain over which one wishes to generalize. Two, whether aggregation is carried out within a single study or over a sample of studies, it is not a panacea. Certain systematic effects, such as sex, race, or attitudes of raters, may bias an entire group of studies (Rosenthal & Rosnow, 1991). Examining large samples of studies through the techniques of meta-analysis (see Chapter 7) is one way of detecting the existence of such variables.

It also seems logical to expect that broader levels of aggregation might be necessary in some situations but not in others. Specifically, Rambo, Chomiak, and Price (1983) examined what Thorndike (1949) labeled extrinsic unreliability and showed that the reliability of performance data is a function both of task complexity and of the constancy of the work environment. These factors, along with the general effectiveness of an incentive system (if one exists), interact to create the conditions that determine the extent to which performance is consistent over time. Rambo et al. (1983) obtained weekly production data over a three-and-a-half-year period from a group of women who were sewing machine operators and a group of women

in folding and packaging jobs. Both groups of operators worked under a piece-rate payment plan. Median correlations in week-to-week (not day-to-day) output rates were sewing = .94; nonsewing = .98. Among weeks separated by one year, they were sewing = .69; nonsewing = .86. Finally, when output in week 1 was correlated with output in week 178, the correlations obtained were still high: sewing = .59; nonsewing = .80. These are extraordinary levels of consistency, indicating that the presence of a production-linked wage incentive, coupled with stable, narrowly routinized work tasks, can result in high levels of consistency in worker productivity. Those individuals who produced much (little) initially also tended to produce much (little) at a later time. More recent results for a sample of foundry chippers and grinders paid under an individual incentive plan over a six-year period were generally consistent with those of the Rambo et al. (1983) study (Vinchur, Schippmann, Smalley, & Rothe, 1991), although there may be considerable variation in long-term reliability as a function of job content.

Challenge #2: Reliability of Job Performance Observation

The issue of reliability of job performance observation is crucial in prediction because all evaluations of performance usually depend on observation of one sort or another, but different methods of observing performance may lead to markedly different conclusions, as was shown by Bray and Campbell (1968). In an attempt to validate assessment center predictions of future sales potential, 78 men were hired as salespeople, regardless of their performance at the assessment center (we discuss the topic of the assessment center in detail in Chapter 13). Predictions then were related to field performance six months later. Field performance was assessed in two ways. In the first method, a trained independent auditor accompanied each man in the field on as many visits as were necessary to determine whether he did or did not meet accepted standards in conducting his sales activities. The field reviewer was unaware of any judgments made of the candidates at the assessment center. In the second method, each individual was rated by his sales supervisor and his trainer from sales training school. Both the supervisor and the trainer also were unaware of the assessment center predictions.

Although assessment center predictions correlated .51 with field performance ratings, there were no significant relationships between assessment center predictions and either supervisors' ratings or trainers' ratings. Additionally, there were no significant relationships between the field performance ratings and the supervisors' or trainers' ratings! The lesson to be drawn from this study is obvious: The study of reliability of performance becomes possible only when the reliability of judging performance is adequate (Ryans & Fredericksen, 1951). Unfortunately, although we know that the problem exists, there is no silver bullet that will improve the reliability of judging performance (Borman & Hallam, 1991). We examine this issue in greater detail, including some promising new approaches, in Chapter 5.

Challenge #3: Dimensionality of Job Performance

Even the most cursory examination of HR practices reveals a great variety of predictors typically in use. Several reviews (Campbell & Wiernik, 2015; Ronan & Prien, 1966, 1971) concluded that the notion of a unidimensional measure of job performance (even for lower level jobs) is unrealistic. Analyses of even single measures of job performance (e.g., attitude toward the company, absenteeism) have shown that they are much more complex than surface appearance would suggest. Despite the problems associated with global criteria, they seem to "work" quite well in most personnel selection situations. However, to the extent that one needs to solve a specific problem (e.g., too many customer complaints about product quality), a more specific criterion is needed. If there is more than one specific problem, then more than one specific criterion is called for (Guion, 1987).

PERFORMANCE AND SITUATIONAL CHARACTERISTICS

Most people would agree readily that individual levels of performance may be affected by conditions surrounding the performance. Yet most research investigations are conducted without regard for possible effects of variables other than those measured by predictors. In this section, therefore, we examine six possible extraindividual influences on performance. Taken together, the discussion of these influences is part of what Cascio and Aguinis (2008b) defined as *in situ* performance: "the specification of the broad range of effects—situational, contextual, strategic, and environmental—that may affect individual, team, or organizational performance" (p. 146). A consideration of *in situ* performance involves context—situational opportunities and constraints that affect the occurrence and meaning of behavior in organizations—as well as functional relationships between variables.

Environmental and Organizational Characteristics

Both absenteeism and turnover have been related to a variety of environmental and organizational characteristics (Allen & Vardaman, 2017; Dineen, Noe, Shaw, Duffy, & Wiethoff, 2007; McEvoy & Cascio, 1987; Sun, Aryee, & Law, 2007). These include organizational factors (e.g., pay and promotion policies, human resources practices); interpersonal factors (e.g., group cohesiveness, friendship opportunities, satisfaction with peers or supervisors); job-related factors (e.g., role clarity, task repetitiveness, autonomy, responsibility); and personal factors (e.g., age, tenure, mood, family size). Shift work is another frequently overlooked variable (Barton, 1994; Staines & Pleck, 1984). Clearly, organizational characteristics can have wide-ranging effects on performance.

Environmental Safety

Injuries and loss of time may also affect job performance (Probst, Brubaker, & Barsotti, 2008). Factors such as a positive safety climate, a high management commitment, and a sound safety communications program that incorporates goal setting and knowledge of results tend to increase safe behavior on the job (Reber & Wallin, 1984) and conservation of scarce resources (cf. Siero, Boon, Kok, & Siero, 1989). These variables can be measured reliably (Zohar, 1980) and can then be related to individual performance. Overall, environmental safety is affected by factors originating at the individual and organizational levels (Hofmann, Burke, & Zohar, 2017).

Lifespace Variables

Lifespace variables measure important conditions that surround the employee both on and off the job. They describe the individual employee's interactions with organizational factors, task demands, supervision, and conditions of the job. Vicino and Bass (1978) used four lifespace variables—task challenge on first job assignment, life stability, supervisor–subordinate personality match, and immediate supervisor's success—to improve predictions of management success at Exxon. The four variables accounted for an additional 22% of the variance in success on the job over and above Exxon's own prediction system based on aptitude and personality measures. The equivalent of a multiple R of .79 was obtained. Other lifespace variables, such as personal orientation, career confidence, cosmopolitan versus local orientation, and job stress, deserve further study (Cooke & Rousseau, 1983; Edwards & Van Harrison, 1993).

Job and Location

Schneider and Mitchel (1980) developed a comprehensive set of six behavioral job functions for the agency manager's job in the life insurance industry. Using 1,282 managers from 50 companies, they examined the relationship of activity in these functions with five factors: origin of the agency (new versus established), type of agency (independent versus company controlled), number of agents, number of supervisors, and tenure of the agency manager. These five situational variables were chosen as correlates of managerial functions on the basis of their traditionally implied impact on managerial behavior in the life insurance industry. The most variance explained in a job function by a weighted composite of the five situational variables was 8.6% (i.e., for the general management function). Thus, over 90% of the variance in the six agency-management functions lies in sources other than the five variables used. Although situational variables have been found to influence managerial job functions *across* technological boundaries, the results of this study suggest that situational characteristics also may influence managerial job functions *within* a particular technology. Performance thus depends not only on job demands but also on other structural and contextual factors such as the policies and practices of particular companies.

Extraindividual Differences and Sales Performance

Cravens and Woodruff (1973) recognized the need to adjust criterion standards for influences beyond a salesperson's control, and they attempted to determine the degree to which these factors explained variations in territory performance. In a multiple regression analysis using dollar volume of sales as the criterion, a curvilinear model yielded a corrected R^2 of .83, with sales experience, average market share, and performance ratings providing the major portion of explained variation. This study is noteworthy because a purer estimate of individual job performance was generated by combining the effects of extraindividual influences (territory workload, market potential, company market share, and advertising effort) with two individual-difference variables (sales experience and rated sales effort).

Leadership

The effects of leadership and situational factors on morale and performance have been well documented (Detert, Treviño, Burris, & Andiappan, 2007; Srivastava, Bartol, & Locke, 2006). These studies, as well as those cited previously, demonstrate that variations in job performance are due to characteristics of individuals (age, sex, job experience, etc.), groups, and organizations (size structure, management behavior, etc.). Until we can begin to partition the total variability in job performance into intraindividual and extraindividual components, we should not expect predictor variables measuring individual differences to correlate appreciably with measures of performance that are influenced by factors not under an individual's control.

STEPS IN CRITERION DEVELOPMENT

Given the previous discussion, we can now describe a five-step procedure for criterion development as outlined by Guion (1961):

1. Analysis of job and/or organizational needs.
2. Development of measures of actual behavior relative to expected behavior as identified in job and need analysis. These measures should supplement objective measures of organizational outcomes such as turnover, absenteeism, and production.

3. Identification of criterion dimensions underlying such measures by factor analysis, cluster analysis, or pattern analysis.
4. Development of reliable measures, each with high construct validity, of the elements so identified.
5. Determination of the predictive validity of each independent variable (predictor) for *each one* of the criterion measures, taking them one at a time.

In step 2, behavior data are distinguished from result-of-behavior data or organizational outcomes and it is recommended that behavior data supplement result-of-behavior data. In step 4, construct-valid measures are advocated. Construct validity is essentially a judgment that a test or other predictive device does, in fact, measure a specified attribute or construct to a significant degree and that it can be used to promote the understanding or prediction of behavior (Landy & Conte, 2016; Messick, 1995). These two poles, **utility** (i.e., in which the researcher attempts to find the highest and therefore most useful validity coefficient) versus **understanding** (in which the researcher advocates construct validity), have formed part of the basis for an enduring controversy in psychology over the relative merits of the two approaches. We examine this controversy in greater detail in the section "Composite Criterion Versus Multiple Criteria."

EVALUATING CRITERIA

How can we evaluate the usefulness of a given criterion? Let's discuss each of three different yardsticks: relevance, sensitivity or discriminability, and practicality.

Relevance

The principal requirement of any criterion is its judged relevance (i.e., it must be logically related to the performance domain in question). As noted in *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organization Psychology, 2018), "A relevant criterion is one that reflects the relative standing of employees with respect to an outcome critical to success in the focal work environment" (p. 14). Hence, it is essential that this domain be described clearly.

Indeed, the American Psychological Association (APA) Task Force on Employment Testing of Minority Groups (1969) specifically emphasized that the most appropriate (i.e., logically relevant) criterion for evaluating tests is a direct measure of the degree of job proficiency developed by an employee after an appropriate period of time on the job (e.g., six months to a year). To be sure, the most relevant criterion measure will not always be the most expedient or the cheapest. A well-designed work sample test or performance management system may require a great deal of ingenuity, effort, and expense to construct (e.g., Jackson, Harris, Ashton, McCarthy, & Tremblay, 2000).

Sensitivity or Discriminability

To be useful, any criterion measure also must be sensitive—that is, capable of discriminating between effective and ineffective employees. Suppose, for example, that quantity of goods produced is used as a criterion measure in a manufacturing operation. Such a criterion frequently is used inappropriately when, because of machine pacing, everyone doing a given job produces about the same number of goods. Under these circumstances, there is little justification for using quantity of goods produced as a performance criterion, since the most effective

workers do not differ appreciably from the least effective workers. Perhaps the amount of scrap or the number of errors made by workers would be a more sensitive indicator of real differences in job performance. Thus, the use of a particular criterion measure is warranted only if it reveals discriminable differences in job performance.

It is important to point out, however, that there is no necessary association between criterion variance and criterion relevance. A criterion element *as measured* may have low variance, but the implications in terms of a different scale of measurement, such as dollars, may be considerable (e.g., the dollar cost of industrial accidents). In other words, the utility to the organization of what a criterion measures may not be reflected in the way that criterion is measured. This highlights the distinction between operational measures and a conceptual formulation of what is important (i.e., has high utility *and* relevance) to the organization (Cascio & Valenzi, 1978).

Practicality

It is important that management be informed thoroughly of the real benefits of using carefully developed criteria. Management may or may not have the expertise to appraise the soundness of a criterion measure or a series of criterion measures, but objections will almost certainly arise if record keeping and data collection for criterion measures become impractical and interfere significantly with ongoing operations. Overzealous HR researchers sometimes view organizations as ongoing laboratories existing solely for their purposes. This should not be construed as an excuse for using inadequate or irrelevant criteria. Clearly a balance must be sought, for the HR department occupies a staff role, assisting through more effective use of human resources those who are concerned directly with achieving the organization's primary goals of profit, growth, and/or service. Keep criterion measurement practical!

CRITERION DEFICIENCY

Criterion measures differ in the extent to which they cover the criterion domain. For example, the job of university professor includes tasks related to teaching, research, and service. If job performance is measured using indicators of teaching and service only, then the measures are deficient because they fail to include an important component of the job. Similarly, if we wish to measure a manager's flexibility, adopting a trait approach only would be deficient because managerial flexibility is a higher order construct that reflects mastery of specific and opposing behaviors in two domains: social/interpersonal and functional/organizational (Kaiser, Lindberg, & Craig, 2007).

The importance of considering criterion deficiency was highlighted by a study examining the economic utility of companywide training programs addressing managerial and sales/technical skills (Morrow, Jarrett, & Rupinski, 1997). The economic utility of training programs may differ not because of differences in the effectiveness of the programs per se, but because the criterion measures may differ in breadth. In other words, the amount of change observed in an employee's performance after she attends a training program will depend on the percentage of job tasks measured by the evaluation criteria. A measure including only a subset of the tasks learned during training will underestimate the value of the training program.

CRITERION CONTAMINATION

When criterion measures are gathered carelessly with no checks on their worth before use either for research purposes or in the development of HR policies, they are often contaminated. Maier (1988) demonstrated this in an evaluation of the aptitude tests used to make

placement decisions about military recruits. The tests were validated against hands-on job performance tests for two Marine Corps jobs: radio repairer and auto mechanic. The job performance tests were administered by sergeants who were experienced in each specialty and who spent most of their time training and supervising junior personnel. The sergeants were not given any training on how to administer and score performance tests. In addition, they received little monitoring during the four months of actual data collection, and only a single administrator was used to evaluate each examinee. The data collected were filled with errors, although subsequent statistical checks and corrections made the data salvageable. Did the "clean" data make a difference in the decisions made? Certainly. The original data yielded validities of .09 and .17 for the two specialties. However, after the data were "cleaned up," the validities rose to .49 and .37, thus changing the interpretation of how valid the aptitude tests actually were.

Criterion contamination occurs when the operational or actual criterion includes variance that is unrelated to the ultimate criterion. Contamination itself may be subdivided into two distinct parts, error and bias (Blum & Naylor, 1968). **Error** by definition is random variation (e.g., due to nonstandardized procedures in testing, individual fluctuations in feelings) and cannot correlate with anything except by chance alone. **Bias,** by contrast, represents systematic criterion contamination, and it can correlate with predictor measures.

Criterion bias is of great concern in HR research and practice because its potential influence is so pervasive. Brogden and Taylor (1950b) offered a concise definition:

> A biasing factor may be defined as any variable, except errors of measurement and sampling error, producing a deviation of obtained criterion scores from a hypothetical "true" criterion score. (p. 161)

Because the direction of the deviation from the true criterion score is not specified, biasing factors may increase, decrease, or leave unchanged the obtained validity coefficient. Biasing factors vary widely in their distortive effect, but primarily this distortion is a function of the degree of their correlation with predictors. The magnitude of such effects must be estimated and their influence controlled either experimentally or statistically. Next, we discuss three important and likely sources of bias.

Bias Due to Knowledge of Predictor Information

One of the most serious contaminants of criterion data, especially when the data are in the form of ratings, is prior knowledge of or exposure to predictor scores. In the selection of executives, for example, the assessment center method (see Chapter 13) is a popular technique. If an individual's immediate superior has access to the prediction of this individual's future potential by the assessment center staff and if at a later date the superior is asked to rate the individual's performance, the supervisor's prior exposure to the assessment center prediction is likely to bias this rating. If the subordinate has been tagged as a "shooting star" by the assessment center staff and the supervisor values that judgment, he or she, too, may rate the subordinate as a "shooting star." If the supervisor views the subordinate as a rival, dislikes him or her for that reason, and wants to impede his or her progress, the assessment center report could serve as a stimulus for a *lower* rating than is deserved. In either case—spuriously high or spuriously low ratings—bias is introduced and gives an unrealistic estimate of the validity of the predictor. Because this type of bias is by definition predictor correlated, it *looks like* the predictor is doing a better job of predicting than it actually is; yet the effect is illusory. The rule of thumb is this: Keep predictor information away from those who must provide criterion data!

Probably the best way to guard against this type of bias is to obtain all criterion data before any predictor data are released. Thus, in attempting to validate assessment center predictions, Bray and Grant (1966) collected data at an experimental assessment center, but these data had no bearing on subsequent promotion decisions. Eight years later the predictions were validated against a criterion of "promoted versus not promoted into middle management." By carefully shielding the predictor information from those who had responsibility for making promotion decisions, a much "cleaner" validity estimate was obtained.

Bias Due to Group Membership

Criterion bias may also result from the fact that individuals belong to certain groups. In fact, sometimes explicit or implicit policies govern the hiring or promotion of these individuals. For example, some organizations tend to hire engineering graduates predominantly (or only) from certain schools. Similarly, we know of an organization that tends to promote people internally who also receive promotions in their military reserve units.

Studies undertaken thereafter that attempt to relate these biographical characteristics to subsequent career success will necessarily be biased. The same effects also will occur when a group sets artificial limits on how much it will produce.

Bias in Ratings

Supervisory ratings, the most frequently employed criteria (Aguinis, 2019; Lent, Aurbach, & Levin, 1971; Murphy & Cleveland, 1995), are susceptible to all the sources of bias in objective indices, as well as to others that are peculiar to subjective judgments (Thorndike, 1920). We discuss this problem in much greater detail in Chapter 5, but, for the present, it is important to emphasize that bias in ratings may be due to spotty or inadequate observation by the rater, unequal opportunity on the part of subordinates to demonstrate proficiency, personal biases or prejudices on the part of the rater, or an inability to distinguish and reliably rate different dimensions of job performance.

COMPOSITE CRITERION VERSUS MULTIPLE CRITERIA

Applied psychologists generally agree that job performance is multidimensional in nature and that adequate measurement of job performance requires multidimensional criteria. The next question is what to do about it. Should we combine the various criterion measures into a composite score, or should we treat each criterion measure separately? If we choose to combine the elements, what rule should we use to do so? As with the issue of utility versus understanding, both sides have had their share of vigorous proponents over the years. Let's consider some of the arguments.

Composite Criterion

The basic contention of Brogden and Taylor (1950a), Thorndike (1949), Toops (1944), and Nagle (1953), the strongest advocates of the composite criterion, is that the criterion should provide a yardstick or overall measure of "success" or "value to the organization" of each individual. Such a single index is indispensable in decision making and individual comparisons, and even if the criterion dimensions are treated separately in validation, they must somehow be combined into a composite when a decision is required. Although the combination of

multiple criteria into a composite is often done subjectively, a quantitative weighting scheme makes objective the importance placed on each criterion used to form the composite.

If a decision is made to form a composite based on several criterion measures, then the question is whether all measures should be given the same weight or not (Bobko, Roth, & Buster, 2007). Consider the possible combination of two measures reflecting customer service, one collected from external customers (i.e., those purchasing the products offered by the organization) and the other from internal customers (i.e., individuals employed in other units within the same organization). Giving these measures equal weight implies that the organization values both external and internal customer service equally. However, the organization may make the strategic decision to form the composite by giving 70% weight to external customer service and 30% weight to internal customer service. This strategic decision is likely to affect the validity coefficients between predictors and criteria. Specifically, Murphy and Shiarella (1997) conducted a computer simulation and found that 34% of the variance in the validity of a battery of selection tests was explained by the way in which measures of task and contextual performance were combined to form a composite performance score. In short, forming a composite requires careful consideration of the relative importance of each criterion measure.

Multiple Criteria

Advocates of multiple criteria contend that measures of demonstrably different variables should not be combined. As Cattell (1957) put it, "Ten men and two bottles of beer cannot be added to give the same total as two men and ten bottles of beer" (p. 11). Consider a study of military recruiters (Pulakos, Borman, & Hough, 1988). In measuring the effectiveness of the recruiters, the researchers found that selling skills, human relations skills, and organizing skills all were important and related to success. They also found, however, that the three dimensions were unrelated to each other—that is, the recruiter with the best selling skills did not necessarily have the best human relations skills or the best organizing skills. Under these conditions, combining the measures leads to a composite that not only is ambiguous but also is psychologically nonsensical. Guion (1961) brought the issue clearly into focus:

> The fallacy of the single criterion lies in its assumption that everything that is to be predicted is related to everything else that is to be predicted—that there is a general factor in all criteria accounting for virtually all of the important variance in behavior at work and its various consequences of value. (p. 145)

Schmidt and Kaplan (1971) subsequently pointed out that combining various criterion elements into a composite does imply that there is a single underlying dimension in job performance, but it does not, in and of itself, imply that this single underlying dimension is behavioral or psychological in nature. A composite criterion may well represent an underlying economic dimension while being essentially meaningless from a behavioral point of view. Thus, Brogden and Taylor (1950a) argued that when all of the criteria are relevant measures of economic variables (dollars and cents), they can be combined into a composite, regardless of their intercorrelations.

Differing Assumptions

As Schmidt and Kaplan (1971) and Binning and Barrett (1989) have noted, the two positions differ in terms of (a) the nature of the underlying constructs represented by the respective criterion measures and (b) what they regard to be the primary purpose of the validation process itself. Let's consider the first set of assumptions. Underpinning the arguments for the

composite criterion is the assumption that the criterion should represent an economic rather than a behavioral construct. The economic orientation is illustrated in Brogden and Taylor's (1950a) "dollar criterion:" "The criterion should measure the overall contribution of the individual to the organization" (p. 139). Brogden and Taylor argued that overall efficiency should be measured in dollar terms by applying cost accounting concepts and procedures to the employee's individual job behaviors: "The criterion problem centers primarily on the quantity, quality, and cost of the finished product" (p. 141).

In contrast, advocates of multiple criteria (Dunnette, 1963a; Pulakos et al., 1988) argued that the criterion should represent a behavioral or psychological construct, one that is behaviorally homogeneous. Pulakos et al. (1988) acknowledged that a composite criterion must be developed when actually making employment decisions, but they also emphasized that such composites are best formed when their components are well understood.

Resolving the Dilemma

Clearly there are numerous possible uses of job performance and program evaluation criteria. In general, they may be used for research purposes or operationally as an aid in managerial decision making. When criteria are used for research purposes, the emphasis is on the psychological understanding of the relationship between various predictors and separate criterion dimensions, where the dimensions themselves are behavioral in nature. When used for managerial decision-making purposes—such as job assignment; promotion; capital budgeting; or evaluation of the cost effectiveness of recruitment, training, or advertising programs—criterion dimensions must be combined into a composite representing overall (economic) worth to the organization.

The resolution of the composite criterion versus multiple criteria dilemma essentially depends on the objectives. Both methods are legitimate for their own purposes. If the goal is increased psychological understanding of predictor–criterion relationships, then the criterion elements are best kept separate. If managerial decision making is the objective, then the criterion elements should be weighted, regardless of their intercorrelations, into a composite representing an economic construct of overall worth to the organization.

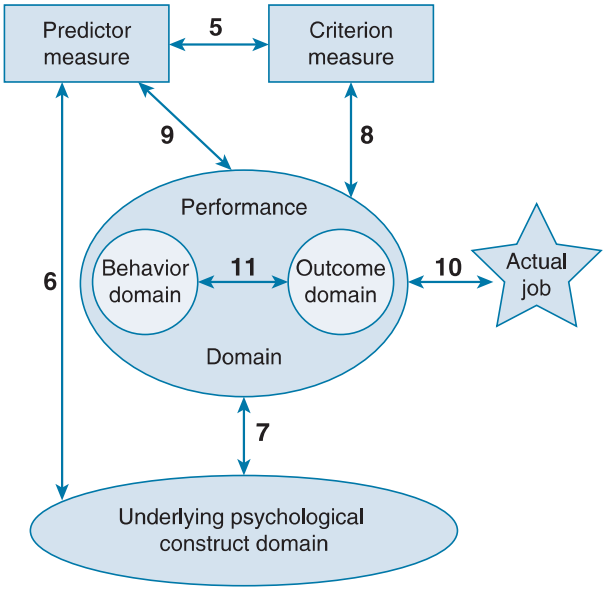
Criterion measures with theoretical relevance should not replace those with practical relevance, but rather should supplement or be used along with them. The goal, therefore, is to enhance utility *and* understanding.

RESEARCH DESIGN AND CRITERION THEORY

Traditionally, personnel psychologists were guided by a simple prediction model that sought to relate performance on one or more predictors with a composite criterion. Implicit intervening variables usually were neglected.

A more complete criterion model that describes the inferences required for the rigorous development of criteria was presented by Binning and Barrett (1989). The model is shown in Figure 4.2. Managers involved in employment decisions are most concerned about the extent to which assessment information will allow accurate predictions about subsequent job performance (Inference 9 in Figure 4.2). One general approach to justifying Inference 9 would be to generate direct empirical evidence that assessment scores relate to valid measurements of job performance. Inference 5 shows this linkage, which traditionally has been the most pragmatic concern to personnel psychologists. Indeed, the term **criterion related** has been used to denote this type of evidence. However, to have complete confidence in Inference 9, Inferences 5 and 8 must be justified. That is, a predictor should be related to an operational criterion

FIGURE 4.2 ■ A Modified Framework That Identifies the Inferences for Criterion Development



Source: Binning, J. F., & Barrett, G. V. Validity of personnel decisions: A conceptual analysis of the inferential and evidential biases. *Journal of Applied Psychology, 74*, 478–494. Copyright © 1989 American Psychological Association.

Note: Linkages in the figure begin with number 5 because earlier figures in the article used numbers 1–4 to show critical linkages in the theory-building process.

measure (Inference 5), and the operational criterion measure should be related to the performance domain it represents (Inference 8).

Performance domains are comprised of behavior–outcome units (Binning & Barrett, 1989). Outcomes (e.g., dollar volume of sales) are valued by an organization, and behaviors (e.g., selling skills) are the means to these valued ends. Thus, behaviors take on different values, depending on the value of the outcomes. This, in turn, implies that optimal description of the performance domain for a given job requires careful and complete representation of valued outcomes and the behaviors that accompany them. As we noted earlier, composite criterion models focus on outcomes, whereas multiple criteria models focus on behaviors. As Figure 4.2 shows, together they form a performance domain. This is why both are necessary and should continue to be used.

Inference 8 represents the process of criterion development. Usually it is justified by rational evidence (in the form of job analysis data) showing that all major behavioral dimensions or job outcomes have been identified and are represented in the operational criterion measure. In fact, work analysis (see Chapter 9) provides the evidential basis for justifying Inferences 7, 8, 10, and 11.

What personnel psychologists have traditionally implied by the term **construct validity** is tied to Inferences 6 and 7. That is, if it can be shown that a test (e.g., of reading comprehension) measures a specific construct (Inference 6), such as reading comprehension, that has been determined to be critical for job performance (Inference 7), then inferences about job performance from test scores (Inference 9) are, by logical implication, justified. Constructs are simply labels for behavioral regularities that underlie behavior sampled by the predictor, and, in the performance domain, by the criterion.

In the context of understanding and validating criteria, Inferences 7, 8, 10, and 11 are critical. Inference 7 is typically justified by claims, based on job analysis, that the constructs underlying performance have been identified. This process is commonly referred to as **deriving job specifications.** Inference 10, by contrast, represents the extent to which actual job demands have been analyzed adequately, resulting in a valid description of the performance domain. This process is commonly referred to as developing a **job description.** Finally, Inference 11 represents the extent to which the links between job behaviors and job outcomes have been verified. Again, job analysis is the process used to discover and to specify these links.

The framework shown in Figure 4.2 helps identify possible locations for what we have referred to as the **criterion problem.** This problem results from a tendency to neglect the development of adequate evidence to support Inferences 7, 8, and 10 and fosters a very shortsighted view of the process of validating criteria. It also leads predictably to two interrelated consequences: (1) the development of criterion measures that are less rigorous psychometrically than are predictor measures and (2) the development of performance criteria that are less deeply or richly embedded in the networks of theoretical relationships that are constructs on the predictor side. These consequences are unfortunate, for they limit the development of theories, the validation of constructs, and the generation of evidence to support important

inferences about people and their behavior at work (Binning & Barrett, 1989). Conversely, the development of evidence to support the important linkages shown in Figure 4.2 will lead to better informed staffing decisions, better career development decisions, and, ultimately, more effective organizations.

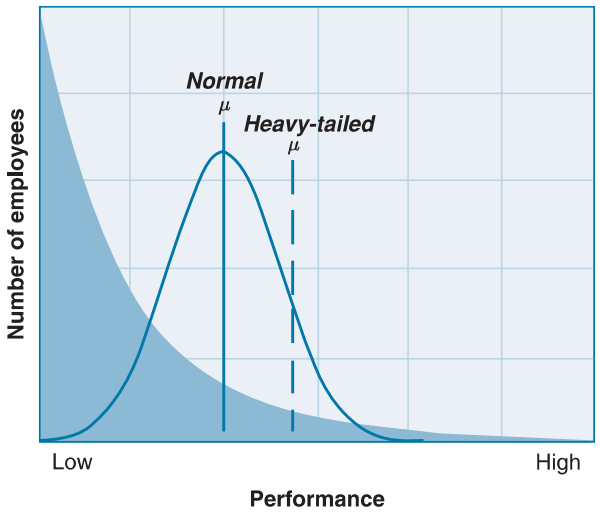
DISTRIBUTION OF PERFORMANCE AND STAR PERFORMERS

Although not referred to explicitly, there is an unspoken assumption that the distribution of criteria, and performance in particular, follows a normal, bell-shaped distribution. For example, as we discuss in detail in Chapter 14, calculations of utility (i.e., financial results of HR interventions) usually make this assumption. If the distribution of performance follows a normal distribution, this means that the majority of individuals are grouped toward the center, and there is a small minority of individuals who are very poor or very good performers (see Figure 4.3).

Challenging this long-standing implicit assumption of normality, a recent stream of research has shown that, for many jobs and occupations, the performance distribution is heavy tailed (Aguinis et al., 2016; Joo, Aguinis, & Bradley, 2017; O'Boyle & Aguinis, 2012). Figure 4.3 includes a visual representation of the assumed normal (i.e., bell-shaped) performance distribution where the majority of scores fall close to the mean μ (i.e., the center of the distribution), with relatively few scores falling at either the low or the high extremes. Figure 4.3 also includes a heavy-tailed distribution, shown in the gray area, also with its own mean.

Figure 4.3 shows a critical difference between these two types of distributions. Specifically, under a heavy-tailed distribution, we expect to see many "star performers" (i.e., those very far to the right of the mean). By contrast, under a normal distribution, the presence of such extreme scores is considered an anomaly. In fact, if one assumes that performance is distributed normally, the presence of such extreme scores is something that needs to be "corrected" through data transformations or outlier-management techniques that could even involve deleting these "abhorrent data points" (Aguinis, Gottfredson, & Joo, 2013b). Also, Figure 4.3 shows that output (i.e., area under the curves) is such that under a heavy-tailed distribution, a minority of individuals are responsible for producing a disproportionate quantity of results. This is not the case in normal distributions.

FIGURE 4.3 ■ Normal and Heavy-Tailed Performance Distributions and Their Means (μ)



Source: Adapted from Aguinis, H., & Bradley, K. J. (2015). The secret sauce for organizational success: Managing and producing star performers. *Organizational Dynamics*, *44*, 161–168.

Do performance distributions follow normal or heavy-tailed distributions? Consider the evidence based on more than 600,000 individual workers, including publications authored by more than 25,000 researchers across more than 50 scientific fields, as well as productivity metrics collected from movie directors, writers, musicians, athletes, bank tellers, call-center employees, grocery checkers, electrical fixture assemblers, and wirers. Results suggest that at least 75% of distributions follow heavy-tailed and not normal distributions (Joo et al., 2017).

Let's illustrate these results more concretely, considering the generic Figure 4.3. If research performance data followed

a normal distribution, there should be approximately 35 researchers with about 10 publications or more each (three standard deviations above the mean). In contrast, results showed that 460 individuals produced that high number of scientific publications. This number is more than 13 times as many as we would expect if the normal distribution were true. Now, consider results for about 3,300 artists who have been nominated for a Grammy award. Five of them would be expected to receive at least 10 nominations under a normal performance distribution. However, 64 artists have received more than 10 nominations. Although this is not true for all jobs and all occupations (Beck et al., 2014; Vancouver, Li, Weinhardt, Steel, & Purl, 2016), it seems that performance distributions are not normal in many, if not most, cases. Consequently, star performers should not be treated as extreme scores or data anomalies that should be "fixed." Rather, the presence of heavy-tailed distributions has a number of important consequences for research in terms of understanding why and when star performers emerge and for practice in terms of how to produce star performers (Aguinis & Bradley, 2015; Aguinis & O'Boyle, 2014):

- It is important to minimize situational constraints (i.e., ceiling constraints) faced by workers to allow for the emergence of star performers. For example, what resources are needed to facilitate the emergence of stars?
- Allow star performers to rotate across teams because this widens their network and takes full advantage of knowledge transfer to rising stars.
- Invest sufficient resources in star performers who are making clear contributions to an organization's core strategic objectives.
- Retain stars by paying attention to their developmental network (e.g., employment opportunities for spouses and long-term contracting with a star's subordinates).
- In times of financial challenges and budget cuts, pay special attention to star performers because once they leave, it will be difficult for an organization to recover. In fact, the departure of a star can create a downward spiral of production when average or even mediocre performers replace stars.
- Give star performers preferential treatment, but clearly articulate these perks to all workers and apply them fairly. In other words, anyone can receive the perks if he or she achieves a high level of performance.
- Invest disproportionate amount of resources into stars, which will likely generate greater overall output and create positive gains.
- The easiest way *not* to produce star performers is to use non-performance-based incentives, encourage limited pay dispersion, and implement longevity-based promotion decisions. Doing so emphasizes homogeneous employee performance.

In conclusion, if the performance distribution is heavy tailed and not a normal shape, then a minority of employees are responsible for a very large and disproportionate amount of results—be it dollars, publications, or Grammy nominations. Obviously, it is in the best interest of organizations not to ignore the presence of star performers but rather to actively attempt to recruit, produce, and retain these stars.

EVIDENCE-BASED IMPLICATIONS FOR PRACTICE

- The effectiveness and future progress of our knowledge of HR interventions depend fundamentally on careful, accurate criterion measurement.
- It is important to conceptualize the job performance domain broadly and to consider job performance as *in situ* performance (i.e., the specification of the broad range of effects—situational, contextual, strategic, and environmental—that may affect individual, team, or organizational performance).
- The notion of criterion relevance requires prior theorizing and development of the dimensions that comprise the domain of performance.
- Organizations must first formulate clear ultimate objectives and then develop appropriate criterion measures that represent economic or behavioral constructs—these could involve behaviors and results. Criterion measures must pass the tests of relevance, sensitivity, and practicality.
- Conclusions reached can depend on (1) the particular criterion measures used, (2) the time of measurement, (3) conditions outside an individual's control, and (4) distortions and biases inherent in the situation or the measuring instrument (human or otherwise).
- Because there may be many paths to success, a broader, richer schematization of job performance is needed.
- Star performers are responsible for a disproportionate quantity of results.

Discussion Questions

1. For what types of jobs is the use of behavior-based performance measures better than results-based measures, and for which other jobs is the reverse true?
2. Discuss the problems that dynamic criteria pose for employment decisions.
3. What are the implications of the typical versus maximum performance distinction for personnel selection?
4. Do you agree with the definition of *counterproductive behaviors*? Are all counterproductive behaviors necessarily bad for organizations—and employees?
5. What are the implications for theory and practice of the concept of *in situ* performance?
6. Are there ethical implications associated with the use of wearable sensors? What would you do if your employer requires that you use a GPS in your car to collect data that can be accessed by the organization 24 hours a day, 365 days a year?
7. How can the reliability of job performance observation be improved?
8. What factors should be considered in assigning differential weights when creating a composite measure of performance?
9. Describe the performance domain of a university professor. Then propose a criterion measure to be used in making promotion decisions. How would you rate this criterion regarding relevance, sensitivity, and practicality?
10. Describe a star performer you know. What are the organizational factors that have allowed this person to thrive?