Validity of Personnel Decisions: A Conceptual Analysis of the Inferential and Evidential Bases

John F. Binning
Illinois State University

Gerald V. Barrett
University of Akron

Issues common to both the process of building psychological theories and validating personnel decisions are examined. Inferences linking psychological constructs and operational measures of constructs are organized into a conceptual framework, and validation is characterized as the process of accumulating various forms of judgmental and empirical evidence to support these inferences. The traditional concepts of construct-, content-, and criterion-related validity are unified within this framework. This unified view of validity is then contrasted with more conventional views (e.g., Uniform Guidelines, 1978), and misconceptions about the validation of employment tests are examined. Next, the process of validating predictor constructs is extended to delineate the critical inferences unique to validating performance criteria. Finally, an agenda for programmatic personnel selection research is described, emphasizing a shift in the behavioral scientist's role in the personnel selection process.

Demonstrating the validity of decisions based on psychological assessment procedures is of fundamental importance to personnel and other applied psychologists. Furthermore, few would argue with the fact that generating and articulating validity evidence is a complex process. To fully appreciate this complexity, it is important to realize that conceptions of validity have evolved over the years through the melding of legal, technical, and practical concerns about the quality and utility of personnel decisions. Inevitably, differences of interpretation and opinion have arisen as each constituency has viewed these myriad concerns from uniquely important perspectives. Perhaps equally inevitable, however, is the confusion that has grown out of these differences. Because this confusion ultimately limits the effectiveness of practitioners and theorists alike, the need for greater clarity cannot be overestimated (Guion, 1987; Landy, 1986; Tenopyr, 1986).

This article is based on the premise that all validity issues discussed in personnel contexts have some conceptual counterpart in the general process of theory development (Landy, 1986). Moreover, various departures from this "ideal" process have led to myopic, if not erroneous, conceptions of validity. To elucidate how these departures have distorted conceptions of validity, the article is divided into four major sections. In the section that immediately follows, we review how the general concept of scientific validity implies a simple model in which constructs and measures of such are inferentially linked. In the next section, we suggest that in personnel selection contexts, a conceptually truncated adaptation of this model often implicitly guides the validation of predictor-criterion relationships. This truncation has for years had an undesirably limiting influence on conceptions of validity. Perhaps its most damaging effect has been the relative neglect of criterion validity concerns. In remedial response to this, a third model is presented. This model is designed to restore and clarify the severed criterion portions of the original. Finally, suggested strategies for elaborating the proposed model and broadening conceptions of validation are discussed.

Validation Vis-à-Vis Theory Development

It is now commonly accepted that validity is not a characteristic of a test or assessment procedure but, instead, of inferences made from test or assessment information (Cronbach, 1970; Guion, 1980, 1987; Landy, 1986; Society for Industrial and Organizational Psychology, 1987; American Psychological Association, 1985). An inference is valid to the extent that it is supported by sound evidence. Expressed alternatively by Nunnally (1978), "one validates not a measuring instrument but rather some use to which a measuring instrument is put" (p. 87). Logically, therefore, to examine the concept of validity in personnel decision making, it is important to delineate (a) the types of inferences involved in applied personnel decision situations and (b) the nature of evidence that can be used to support such inferences.

Inferences Linking Psychological Constructs

Following Landy's (1986) lead, it is both appropriate and important to view the process of validating a particular selection procedure as a special case of hypothesis testing and scientific
theory building. The following rudimentary characterization of the theory-building process will provide a backdrop for further discussion of some important validity concepts.

Psychological constructs are labels for clusters of covarying behaviors. In this way, a virtually infinite number of behaviors is reduced to a system of fewer labels, which simplifies and economizes the exchange of information and facilitates the process of discovering behavioral regularities. For example, it is less cumbersome to refer to the relation between verbal and quantitative ability than to the abilities to add, subtract, multiply, and divide numbers, fractions, decimals, and so forth, and their relations to reading, spelling, understanding word meanings, and so on.

Putting aside the perennial debate over the objective existence of psychological traits and psychologists' constructs (Cronbach & Meehl, 1955; Kane, 1982; Loevinger, 1957; Messick, 1981; Nunnally, 1978), viewed pragmatically, a construct is merely a hypothesis about which behaviors will reliably covary. Constructs are heuristic devices for describing behavioral domains. Of course, construct domains can vary in being large versus small, specific versus general, and fuzzy versus clearly defined (Guion, 1987; Nunnally, 1978). Also, constructs become the object of conceptual scrutiny in their own right. In other words, psychologists hypothesize both (a) whether certain behaviors will covary and (b) whether the clusters of covarying behaviors (constructs) tend to covary in meaningful ways. In this general sense, the terms construct validation and theory development imply the same basic process. Both refer to the process of identifying (and often reifying) constructs by developing measures of such constructs and examining relationships among the various measures.

Nunnally (1978) delineated the four inferences that form the core of this construct validation process. These four inferences logically bind the components of the model presented in Figure 1. One can attempt to determine whether an inferred relationship between two constructs (e.g., anxiety and manual dexterity) exists by developing measures or causal conditions for each (labeled X and Y, respectively). It is important to emphasize that these measures are nothing more than procedures for sampling behaviors within the respective construct domains. The following four inferences then follow logically:

1. X and Y relate in some specified way.
2. X is a measure of (or treatment that induces) anxiety.
3. Anxiety and manual dexterity are causally related in some specified way.
4. Y is a measure of (or treatment that induces) manual dexterity.

Even though these four inferences are interrelated, a single experiment cannot validate all four inferences simultaneously. In fact, Inference 1 is the only one that can be empirically tested directly. That is, we can use our measures of anxiety and manual dexterity to derive scores that are subsequently found to relate either experimentally or correlationally. These data serve as empirical evidence of the veridicality of Inference 1. From this one empirical finding, therefore, it would be necessary to infer the truth or falsity of the others, because Inferences 2, 3, and 4 each link an observable measure with a hypothetical construction. Of course, merely finding a correlation between X and Y leaves open several alternative interpretations of possible relationships. For example, anxiety and manual dexterity are perhaps both related to some third construct?

To provide incontrovertible proof that the four inferences are correct, it would be necessary to empirically demonstrate three of the inferences. If three of the linkages are unequivocally proven correct, then complete confidence in the fourth would be justified. However, because this direct empirical proof is impossible (Nunnally, 1978), typical practice is to assume that two of the three inferences (2, 3, or 4) are correct and this, combined with empirical evidence of Inference 1, allows a valid conclusion regarding the remaining inference. Generally, these conclusions about construct validity are strengthened in those situations in which the truth of the assumptions is obvious to everyone scrutinizing the conclusions drawn. Specifically, we are more confident that a test validly measures a given construct if (a) the behavioral domain of the other construct is explicitly defined and (b) the assumption of a relationship between the two constructs is unarguable (Nunnally, 1978).

The Three Faces of Construct Validity

To avoid confusion, it is important to realize that the term construct validity has thus far been used to describe the soundness of evidence supporting any of the four inferences. Thus, the term is being used in its most general sense in reference to construct–construct links (Inference 3), construct–measure links (Inferences 2 and 4), and measure–measure links (Inference 1). However, what have traditionally been of particular concern to research psychologists and psychometrists are construct–measure links (i.e., Inference 2 or 4). In the heyday of trait psychology, construct validity often referred to whether a given test or measurement procedure allowed accurate inferences about an individual's standing on a psychological construct of particular interest (D. T. Campbell & Fiske, 1959; Cronbach & Meehl, 1955; Ebel, 1977; Guion, 1980; Messick, 1980). These two uses of the term construct validity (equal concern for Inferences 1, 2, 3, & 4 vs. primary concern for only Inference 2 or 4) are clearly congruent. In fact, the difference in perspective was recognized by Loevinger (1957) when she referred to the validity of the construct versus the validity of the

![Figure 1. Critical inferential linkages in the theory-building process.](image-url)
test as a measure of the construct (Landy, 1986). If theory building is of primary interest, Inferences 1, 2, 3, and 4 are all of equal importance. On the other hand, in specific situations (e.g., development of a new test), Inference 2 (or 4) is emphasized. This becomes potentially more confusing because the term construct validity has a somewhat different connotation in the personnel selection literature. Here, it has been frequently used to describe a specific evidential approach for justifying a specific measure–construct link (i.e., the predictor–performance linkage portrayed in Figure 2) by documenting underlying construct–construct and construct–measure links (Schwab, 1980). The inferences implicated in this latter meaning are described in detail in the next section. Perhaps the most important issue at this juncture is to realize that these various meanings of the term construct validity are nothing more than different views of the same logical system, with varying emphasis on different inferences.

Examining Traditional Conceptions of Validity

A common conception of the personnel selection process involves (a) analysis of the job to determine (b) a performance domain, defined in terms of job behaviors or outcomes, which then guides (c) the selection or development of certain assessment procedures, which make possible (d) predictions about the likelihood that applicants will perform the job with a certain degree of proficiency, and then subsequently (e) evaluating individual performance by some operational criterion measure (e.g., Cascio, 1987; Muchinsky, 1987; Society for I/O Psychology, 1987; APA, 1985; Uniform Guidelines, 1978). This process implies a framework, presented in Figure 2, which parallels Figure 1 in many respects. The framework represented in Figure 2 portrays the following inferences:

5. Predictor measurements relate to criterion measurements.
6. The predictor measure is an adequate sample from a psychological construct domain.
7. The predictor construct domain overlaps with the performance domain.
8. The criterion measure is an adequate sample from the performance domain.
9. The predictor measure is related to the performance domain.

These inferences serve to link the components in Figure 2 analogously to the inferences in Figure 1. It is important to realize, however, that in the transition from Figure 1 to Figure 2, two important differences have arisen. First, an additional measure–construct linkage (Inference 9) has been created, linking the predictor measure and the performance domain. Second, rather than equal emphasis being placed on all inferences, this additional measure–construct (Inference 9) link has taken on greater relative importance.

The implications of this way of thinking for understanding the validation process are explored in the discussion that follows. Before detailing how validation of personnel selection decisions is merely a special case of the more general construct validation process, it would be helpful to discuss the process of conceptualizing and constructing behavioral domains.

Contrasting Predictor Construct Domains and Performance Domains

In an attempt to simplify the virtually infinite number of behaviors that can be exhibited by human beings, psychologists attempt to identify naturally occurring clusters, then construct labels for them, and investigate the covariance between them. Predictor constructs, therefore, represent those clusters of covariant behaviors identified through psychological research and constructed to enhance our general understanding of behavior.

In contrast to the psychologists' search for naturally occurring behavioral construct domains across myriad situations, organizational designers in effect create behavioral domains to enhance their understanding and prediction of job behavior. In fact, it is important to realize that from our pragmatic viewpoint, a job performance domain is a construct, albeit in a conceptually different sense than is usually implied in the psychological literature. Nonetheless, the performance of any job in any organization is a cluster of interlocked and covariant behaviors, and this cluster consists of a subset of all possible behaviors necessary for the organization to accomplish its broader goals and objectives (Weick, 1979). Just as psychological constructs represent behavioral domains, performance associated with a job (or distinguishable aspects of job performance) represents a behavioral domain.

Performance domains are conceptually distinct from predictor constructs in that the universe to be sampled is delineated differently. Construct domains on the predictor side are conceived of by the research psychologist with reference to some theoretical framework developed to explain general regularities in human behavior. Performance domains are determined, or at least influenced, by organizational decision makers and selection specialists collaborating to translate broad organizational objectives into normative statements of valued behaviors and outcomes.

The overriding reason for constructing behavioral domains on both the predictor and the performance side is the parceling of myriad behaviors into meaningful clusters to enhance understanding and communication. However, this parceling process is different on the predictor versus the performance side because of differences in (a) the conceptualization of predictor domains versus performance domains, and (b) the specific purposes for speci-
fying behavioral domains, (c) the methods used to cluster behaviors, and often (d) the language system used to communicate about the resulting construct systems.

First, the source of covariance between job behaviors is designed by the organization and induced by various external control and coordination mechanisms (Mintzberg, 1983). This can be contrasted with the naturally occurring covariance resulting from the individual’s personal predispositions or the interaction of these predispositions with untold environmental influences, as is often conceptualized for psychological constructs. Second, predictor constructs are clusters of behaviors created by research psychologists to capture general regularities in behavior. Performance domains are designed more or less rationally to interlock in such a way as to maximize efficient attainment of organizational goals. As a result, performance domains are clusters of behavior–outcome units that are differentially valued by the organization. Depending on which goals are most operative, clustering systems can vary considerably (Campion & Thayer, 1985; Griffin, 1982; Harvey, 1986).

Third, through psychologists’ use of both correlational (e.g., factor analytic) and experimental methods, behavior is empirically examined, and reliable regularities are conceptualized and assigned construct labels. In contrast, organizational designers typically rely on rational and relatively informal methods for delineating performance domains. Finally, construct terminology on the predictor side reflects the concern for identifying domains of behavior caused by personal dispositions (e.g., “she is extraverted”). Terminology used in organizations to describe performance domains is more often goal-related (e.g., “she is customer-service oriented”). These different terminologies for describing behavioral domains have been discussed extensively elsewhere (Fleishman, 1982; Pearlman, 1980).

On the predictor side, behaviors can be clustered hierarchically in varying levels of inclusiveness. For example, Dunnette (1976) reviewed attempts to conceptualize intellectual functioning by pointing out that behavior can be grouped into a single, global construct (i.e., the g factor), several less inclusive constructs (e.g., Thurstone’s, 1938, seven primary mental abilities), or literally hundreds of constructs (e.g., Guilford’s, 1967, structure of intellect model).

Organizations also conceptualize job behavior at different levels of inclusiveness, depending on the purpose at hand. For purposes of administrative decision making, global “constructions” of job behavior are evoked, because the overriding imperative is the comparison of employees’ overall contribution to the organization. In this situation, the constructed system of performance domains can be conceptualized as merely a system of different job titles, each connoting a different domain of behaviors. On the other hand, when remedial feedback about job behavior is required, organizations often cluster performance into a number of behaviorally meaningful dimensions. For example, the process of constructing behaviorally anchored rating scales (Bernardin & P. C. Smith, 1981; P. C. Smith & Kendall, 1963) can be viewed as resulting in a performance construct system that enhances intraorganizational communication and decision making about job performance in a specific organization (Feldman, 1986). A source of difficulty for many job analysts and performance appraisal system designers is the fact that the organization’s conceptual system for describing performance is ineffective for certain purposes.

Performance domains result from the division of labor fundamental to organizing human activity. The conceptualization and resulting terminology used in an organization to describe performance differences (both within and between jobs) serves to make behavior in organizations more understandable and orderly. Building on Weick’s (1979) basic tenet that organizing is a “consensually validated grammar for reducing equivocality by means of sensible interlocked behavior” (p. 3), we propose that performance construct systems are an important part of the grammar and culture of a given organization. The manner in which performance behavior is clustered and labeled is part of the consensually validated conceptual scheme that helps make sense out of the complex stream of interlocked behaviors in the organization.

Viewed from this perspective, selection decisions represent attempts to identify regularities in applicants’ behavior, but only those behaviors identified by the organization as valuable for coordination with others’ behavior that are necessary for goal attainment. Personnel selection, then, is the process of identifying and mapping predictor samples of behavior to effectively overlap with performance domains. Validity, therefore, can be viewed as the extent to which these two construct systems overlap.

The “Unitarian” Conception of Validity

The trilogy of construct, content, and criterion-related validities was first articulated in the “Technical Recommendations for Psychological Tests and Diagnostic Techniques” (American Psychological Association, American Educational Research Association, & National Council of Measurement Used in Education, 1954). As noted by Landy (1986), this trilogy was quite valuable in that it enhanced the clarity with which validity concepts were typically discussed at the time. As is the case with many popular conceptualizations, however, its initial usefulness was replaced by growing confusion. This confusion is due, in part, to the tendency for certain erroneous interpretations, misconceptions, and legal mandates to become crystallized as part of professional psychology’s conventional wisdom (G. V. Barrett, 1972) or tenets of orthodoxy (Guion, 1976). It was many years before this conventional wisdom was questioned in a systematic way (Dunnette & Borman, 1979; Guion, 1977, 1978, 1980; Messick, 1975, 1980; Tenopyr, 1977; Tenopyr & Oeljen, 1982) and, by then, confusion was running rampant.

For many years, the concepts of construct, content, and criterion-related validity have been described as different types of validity. Some recent descriptions have gone so far as to suggest that each of these validity analysis strategies (Lawshe, 1985) should be chosen according to the kinds of inferences or conclusions one wishes to make about job applicants (e.g., Lawshe, 1985; Saal & Knight, 1988) or the nature of the selection procedure (e.g., R. S. Barrett, 1980). Although the latter view has to some degree been induced by the prevailing opinion in Title VII litigation, this linking of different validities to different inferences or types of predictors is logically problematic because of the implication that in any given decision situation, only one of
the three validity concepts is useful. On the contrary, an inference drawn from currently available information about some aspect of future job performance (Inference 9) is the single overriding inference; and content-, construct-, and criterion-related considerations are all quite relevant for justifying its validity. These three concepts are more appropriately viewed as labels for three evidential bases (Messick, 1980) from which inferences about future job performance can be supported or justified.

The applied decision maker is concerned about the extent to which test or assessment information will allow accurate predictions about subsequent job performance (Inference 9). One general approach to justifying Inference 9 would be to generate direct empirical evidence that assessment scores relate to valid measurements of job performance (Sussmann & Robertson, 1986). Inference 5 represents this linkage, which has historically been of primary pragmatic concern to personnel psychologists. The term criterion-related has traditionally been used to denote this type of evidence and, in fact, often implies the unnecessary restriction that only correlational evidence is appropriate. As Landy (1986) ably pointed out, substantive theories are seldom, if ever, built solely on correlational evidence. Viewed in this way, criterion-related evidence can be experimental and quasi-experimental in nature.

Why, therefore, have personnel specialists relied so heavily on correlational evidence of validity? This bias might derive from the fact that in personnel selection situations, the constructs of interest are conceived of as enduring person characteristics (e.g., abilities) on the predictor side and fixed job performance measures on the criterion side. Neither of these is typically thought to be amenable to experimental manipulation, except under the most contrived laboratory conditions. Perhaps another factor contributing to this bias in favor of correlational evidence was the conventionally held belief in the situational specificity of validity (Schmidt & Hunter, 1981), which would preclude the use of laboratory analogues to real work settings.

Logically, then, one approach for justifying Inference 9 is to empirically link the predictor and the criterion. However, this results in only partial justification. Analogous to the validation of inferences in Figure 1, to have complete confidence in the validity of Inference 9, both Inferences 5 and 8 must be justified. The relative neglect of Inference 8 by those collecting criterion-related evidence represents a critical truncation of the validation process. Suffice it to say that for criterion-related evidence to be a compelling argument for Inference 9, strong evidence of both Inferences 5 and 8 is required.

What personnel specialists have traditionally implied by the label construct validity is tied to Inferences 6 and 7. Analogous to the logic presented earlier, it is assumed that if Inferences 6 and 7 can be supported by sound evidence, then one can confidently believe Inference 9 to be true. The difference is merely one of focus. Therefore, the general conception of construct validity is merely viewed differently in the context of validating personnel selection decisions. In a selection context, Inference 9 is most critical. If it can be shown that a test measures a specific construct (Inference 6) 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.

How does a personnel selection specialist support Inferences 6 and 7? Evidence supporting Inference 6 primarily takes the form of empirically based relationships and judgments that are both convergent and discriminant in nature (D. T. Campbell & Fiske, 1959; Cook & D. T. Campbell, 1979; Cronbach & Meehl, 1955; Dragow & Miller, 1982; Rezmovic & Rezmovic, 1981). Convergent evidence exists when (a) test scores relate to scores on other tests of the same construct, (b) test scores from people who differ in the extent to which they possess the focal construct also differ in a predictable way, or (c) test scores relate to scores on tests of other constructs that are theoretically expected to be related. Discriminant evidence occurs when test scores do not relate to scores on tests of theoretically independent constructs. Note that this discussion can apply equally to criterion measurement (Inference 8).

Inference 7, because it links two hypothetical behavioral domains, cannot be examined empirically. Analogous to Inference 3 in Figure 1, Inference 7 must be justified theoretically and logically on the basis of accumulated knowledge of construct-construct relations. On closer conceptual scrutiny, however, the analogy loses its relevance because the two constructs being related do not share common nomological (Margenau, 1950) status. The unique conceptual issues that arise when relating predictor and performance domains are examined in detail later. For now, to the extent that Inferences 6 and 7 are supported, the use of the predictor test to predict job performance is construct valid.

A third approach for justifying Inference 9 involves demonstrating that the predictor is isomorphic and obviously interchangeable with the performance domain. This line of reasoning is particularly defensible when it is realized that predictor tests are always samples of behavior from which we infer something about behavior on a job (Dunnette, 1963). The behaviors sampled may be dissimilar or similar ("sign vs. sample") to the criterion behaviors being predicted (Wernimont & Campbell, 1968). If an applicant performs behaviors as part of the assessment phase that closely resemble behaviors in the performance domain, then many personnel specialists feel that, logically, the inference about future job performance is better justified. This line of reasoning underlies the type of evidence traditionally labeled content validity. Of course, various specific procedures for analyzing the degree of isomorphism between predictors and criteria have been proposed (Doran, 1987; Foley & Sundstrom, 1985; Hamilton, 1981; Lawshe, 1975; Schmitt & Ostroff, 1986; Trattner, 1982), but the same basic logic underlies each.

Content-related evidence of validity has traditionally involved justifying Inference 9 by rational examination of the manner in which the performance domain is sampled by the predictor. Analogous to statistical sampling theory, if a predictor sample is constructed in congruence with certain principles (e.g., ensuring representativeness as well as relevance of the sample), one can assume that scores from that sample will accurately estimate the universe from which the sample is drawn. It is this emphasis on operationalization and sample construction that motivated Tenopyr (1977) to refer to content validation as "content-oriented test construction" (p. 52). Therefore, when a
Decision Validity Versus Predictor Development

Thus far, the concepts of construct-, content-, and criterion-related evidence have been discussed solely as evidential bases for justifying decision validity. However, the implications of differences between the three can be traced back in the decision-making process. By doing so, their differences can be more clearly appreciated.

Personnel decision making involves two fundamental phases: (a) constructing the predictor as a sample of some behavioral domain and (b) using this behavioral information to make predictions about future job behavior. This latter data combination phase is the immediate precursor to employment decisions and has therefore received considerable legal and professional scrutiny. Yet, the data collection phase, which involves specifying the behavioral data base, has equally important implications for subsequent decision quality (Sawyer, 1966). The respective roles of the construct-, content-, and criterion-related concepts in the development of predictor samples of behavior deserves conceptual scrutiny.

With reference to Figure 2, the point of departure for the development of any personnel selection system is the performance domain. From this delineation of desirable job behaviors or outcomes, selection specialists “work backwards” to specify which behaviors or outcomes should be sampled by the predictors. There are three routes from the performance domain to predictor development: The construct-related approach involves identifying psychological construct domains that overlap significantly with the performance domain (Inference 7) and then developing predictors that adequately sample these construct domains (Inference 6). The content-related approach involves developing predictors that directly sample the performance domain. The criterion-related approach involves developing some operational measure of behaviors in the performance domain (Inference 8) and then identifying or developing predictors that will relate empirically with the operational criterion measure (Inference 5).

We would like to draw attention to a fundamental difference between the criterion-related approach and the other two approaches. The criterion is merely an operational sample of the performance domain. At its best—that is, being neither deficient nor contaminated—it taps the entire performance domain, and the criterion-related approach reduces logically to the content-related approach. At its worst, it represents an atheoretical and circuitous, if not an entirely misleading route, to predictor development (e.g., “dust-bowl empiricism”). From this perspective, we propose that the construct-related and content-related approaches represent the two fundamental predictor sampling strategies. Construct-related implies that predictor sampling is guided by evoking a psychological construct domain. Content-related implies that predictor sampling is guided by evoking a performance domain. To the extent that the two domains are derived differently and relations between the two are not well understood, construct- and content-related approaches can lead to substantive differences in predictor development and consequent decision validity (R. S. Barrett, 1980).

In contrast with the construct- and content-related approaches, the criterion-related approach is best characterized as a research strategy for empirically assessing the quality of either predictor sampling strategy. Viewed from this perspective, judgments of validity are tantamount to judgments about the adequacy of behavior sampling (construct- and content-related) or empirical indexes of such adequacy (criterion-related).

Generating Evidence for Decision Validity

There has been considerable debate over the years regarding whether the construct-related versus the content-related viewpoint provides the most fruitful model for guiding predictor development and subsequent decision making. For instance, a fundamental conceptual issue was raised when Wernimont and J. P. Campbell (1968) argued that the classic validity model and its emphasis on predictor tests as signs of underlying constructs should be replaced by the behavioral consistency approach in which predictors represent samples of job behavior. Upon closer examination, this issue is really one of how predictor domains should best be specified. If the predictor test is labeled a sign, it implies that the behavior domain was specified by the theory surrounding a psychological construct. If the predictor test is labeled a sample, it implies that the behavior domain was specified by the “theory” surrounding job performance. Ultimately, the resolution of this controversy depends on one’s realizing that the two approaches are inextricably intertwined in the inferential system portrayed in Figure 2. Interestingly, this distinction parallels in certain respects the long-standing controversy in personality psychology between traditional trait versus situationalist approaches for predicting behavior (Mischel, 1973). The issues of whether intrinsic attributes versus environmental characteristics are the most potent influences of behavior are certainly not fully resolved; however, advances have been made to integrate the trait–situation perspectives (Kenrick & Funder, 1988; Mischel, 1984; Schneider, 1987).

In our view, personnel psychologists should never avidly recommend the abandonment of construct-based theory development, because it is the hallmark of fruitful scientific inquiry. Tenopyr (1977) pointed out that for a test to have high predictive value, it must share the same psychological constructs that underlie job behavior. This view recognizes that content specification is part of the construct validation process. That is, part of justifying that a test measures a given construct is the examination of the internal structure of the test to assess the extent to which it is consistent with the theory surrounding the construct. Irrespective of this conceptual unity, the construct versus content perspectives are explicitly recognized in the Uniform Guidelines (1978), Standards (1985), and Principles (1987), and therefore it is pragmatically important to draw clear operational and semantic distinctions between them.

Criterion-related evidence is by its nature empirical, whereas content-related and construct-related evidence are typically conceived of as relying more on human judgment and thus are used differently to justify inferences from test scores. Perhaps
because the precision introduced by careful quantification of psychological phenomena is fundamental to scientific inquiry, criterion-related evidence has been endorsed as legally superior to the other two forms of evidence (Uniform Guidelines, 1978). However, the scientific superiority of criterion-related evidence has not received this endorsement in professional guidelines (e.g., Society for I/O Psychology, 1987; APA, 1985).

Although this issue has been debated for years, the present framework makes it clear that there is no inherent or immutable superiority of criterion-related evidence (especially when restricted in form to predictor–criterion correlations) over other lines of evidence. An uncritical bias in favor of criterion-related evidence can have deleterious effects on theoretical understanding. Validation research for assessment centers provides a case in point. After reviewing the empirical validity evidence, Klimoski and Brickner (1987) concluded that despite consistent empirical evidence, the theoretical explanations of "the predictive validity of assessment centers remains a puzzle" (p. 256). G. V. Barrett, Alexander, O'Connor, and Forbes (1978) argued that coincidental empirical relationships can be discovered when relying on a "dust-bowl empiricism" approach to validation. These coincidental relationships, when atheoretically discovered and interpreted, will detract from our ultimate understanding of complex criterion behavior, because they lack what Guion (1980) has called job-relatedness. The work of Schmidt and Hunter (1981) also casts suspicion on the reliability of evidence from individual criterion-related validity studies due to the excessive sampling error that results from the use of small validation samples.

One could reasonably argue that content-related and construct-related evidence, when based on sound professional judgment about appropriate test use, are often superior to criterion-related evidence. Research does indicate that pooled estimates of criterion-related validity, based on the opinions of personnel psychologists, are more accurate than empirical evidence obtained from small-sample validity studies (Hirsch, Schmidt, & Hunter, 1986; Schmidt, Hunter, Croll, & McKenzie, 1983). The traditional emphasis placed on criterion-related evidence may suggest only that evidence largely based on judgment is more likely to be questioned because of the widely held belief that judgments are inherently fallible. It may also suggest that people do not fully realize the subjective nature of judgments about the relevance of criterion measures.

In some discussions of validity, an appeal is made to follow "professionally accepted procedures" for generating evidence (Lawshe, 1985). It is easily argued that there is nothing approaching a specific, unambiguous set of professionally accepted standards for determining the validity of inferences from test scores (Landy, 1986). Even casual examination of legal testimony and professional literature indicates that no consensus exists regarding issues of content-related validation (Kleiman & Foley, 1978); concurrent versus predictive validation (G. V. Barrett, Phillips, & Alexander, 1981; Guion & Cranny, 1982; Schmitt, Gooding, Noe, & Kirsch, 1984); adequate sample sizes (Monahan & Muchinsky, 1983; Schmidt, Hunter, & Urry, 1976); validity generalization (Burke, 1984; Callender & Osburn, 1980; Gutenberg, Arvay, Osburn, & Jeanneret, 1983; James, Demaree, & Mulaik, 1986; Schmitt & Noe, 1986); and criterion development (G. V. Barrett & Kernan, 1987; Kleiman & Durham, 1981), to mention only a few. As a matter of fact, the validation of certain inferences is actually being made by the courts rather than by professionals who regularly use the measuring instruments upon which these inferences are based. For instance, from a breathalyzer reading of .10, the courts infer that driving ability is impaired. This may not be a valid inference, even though the test, by all psychometric and professional standards, may be valid for measuring blood alcohol levels. A number of the same issues will likely arise regarding the use of polygraphs, drug tests, and genetic screening. Similarly, some of the early courts attempted to mandate that a certain validity coefficient, per se, made a test valid or not. Clearly, the process of drawing inferences from test scores is a very complex one, particularly if one considers the interrelated roles of technical, practical, and legal opinion.

**Which Inference Is Which?**

Throughout this article, the goal has been to delineate the inferences that logically underlie the process of validating psychological constructs and measurement procedures. However, there is another sense in which multiple inferences are discussed by personnel decision makers. There are many potential inferences about future job behaviors that may be drawn from the same test scores. These multiple inferences about future job behaviors should not be confused with the inferences represented in Figures 1 and 2. The many different inferences about future job behavior are merely specific examples of Inference 9 in Figure 2. It is possible to conceive of tests that yield scores that are unquestionably valid indicators of some underlying, theoretically meaningful construct (see Ebel's, 1961, discussion of measurement in the physical sciences). Of course, what is valid is the inference that test scores reflect differences in the construct (e.g., Inference 6), but this is conceptually quite different from inferences about future job behavior drawn from test scores (Inference 9). Guion (1974, 1987) highlighted this distinction by differentiating between job relevance and validity of trait measurement. With enough theoretical and empirical corroboration, it can be confidently concluded that test scores and construct differences co-vary systematically. Therefore, test scores make valid inferences about the construct possible. Yet, one may attempt to infer whether a person who scores in a particular way on the test will perform in a certain way on a job, training program, and so on, but these are quite different kinds of inferences. The test may not be valid for some or all of these purposes, because each implies a different criterion to be predicted. Likewise, the inferences about job performance may be valid, but inferences about other outcomes (e.g., tenure) may not.

It is important to make this distinction to emphasize that a test can be construct valid (in the sense that it validly measures a given construct) and yet certain inferences about future behavior may not be valid. For selection purposes, then, this test would not be construct valid in the traditional Title VII sense. It is this differential past usage of the term construct validity that motivated Guion (1980) to "identify the unifying concept of validity as similar, but not necessarily identical, to what has
when discussing validity of the predictor are equally important in the unifying sense to refer to the justifiable confidence in our selection decision. Construct-related validity should be reserved for references to a particular evidential approach to demonstrating validity that focuses on justifying certain critical construct–measure and construct–construct inferences (Inferences 6 and 7).

Construct-Related Validity of the Criterion

The discussion thus far has reflected a common view of validation. That is, personnel specialists generally place more emphasis on validity of the predictors because the overriding organizational imperative is to gather predictor information as the basis for important and inevitable selection decisions. However, we would like to call attention to Cascio's (1987) statement that "in order to emerge from the 'dark ages' we need clear thinking, in-depth theorizing about criteria, and identification of the goals of criterion measurement" (p. 51). The importance of this statement should become clearer on examination of Figure 3, which represents an adaptation of the validation paradigm presented in Figure 2. The model presented in Figure 3 was designed specifically to link several traditional concepts unique to personnel decision making and to highlight the conceptual differences between predictor construct and performance domains.

Note that the systems of inferences detailed in Figures 1, 2, and 3 are logically symmetrical. Therefore, the issues raised when discussing validity of the predictor are equally important for validating criterion measures (Frederiksen, 1986). The caveat proposed here is that criterion measures must be validated analogously to predictors (Guion, 1961, 1976, 1987; James, 1973), with reference to the inferential linkages being supported by evidence. The importance of this point is often underestimated.

In a typical selection situation, again, Inference 9 is the critical inference for which confirming evidence is required. The validation process involves accumulating evidence of various forms to justify Inference 9 in either a direct empirical way (e.g., validity coefficients, contrasted groups, or test construction analyses) or more judgmentally by confirmation of Inferences 6, 7, and 8. Inference 7 represents whether a specific psychological construct underlies job performance, whereas Inference 8 represents whether the operational criterion samples the performance domain. Generating evidence for Inferences 7 and 8 is the process of accumulating construct-related evidence of criterion validity.

The present framework helps to identify possible loci for the criterion problem. It results from a tendency to truncate the nomological network (specifically, Inferences 7, 8, and 10), which in turn leads to a myopic view of criterion validity. Two interrelated effects of this myopia are likely to result. First, the development of criterion measures is likely to be less psychometrically rigorous than predictor development. Wiggins (1973) stated that "basically, the 'problem' resides in the considerable discrepancy that typically exists between our intuitive standards of what criteria of performance should entail and the measures that are currently employed for evaluating such criteria" (p. 39). Second, performance criteria are likely to be less deeply or richly embedded in networks of theoretical relationships than are constructs on the predictor side. Perhaps this state of affairs has resulted partially from the differences between research for administrative prediction and research for scientific understanding (Anderson & Shanteau, 1977; Loevinger, 1957). Research for prediction tends to ignore the importance of multideterminant functional relationships between variables.

The value of an employee's behavior or accomplishments to an organization is ultimately a relative value judgment by some member or members of the organization (Fiske, 1951). Stated strongly, Campbell (1983) maintains that "the meaning of performance is not something to be 'discovered'; it should be imposed" (p. 286). As such, it is amenable to different interpretations depending on who is making the judgment. As dominant coalitions or critical alliances (Weick, 1979) shift and the organization's values change, so do normative judgments of an employee's worth (Guion, 1961). As a result, it is less likely that the systematic procedures that characterize professional test development will be applied to criterion development (Banks & Roberson, 1985). Once a test is rigorously developed, it has perceived potential for long-term usefulness in various prediction and assessment applications. The same kind of rigor in the development of criterion measures, even assuming that the organization would "foot the bill," might quickly lose its utility with a change in values that often occurs with a change in the organization's leadership. Also, idiosyncratic values regarding behavior in different organizations logically require customized criterion measurement systems. Still another factor contributing to a lack of substantive criterion development is the long-held belief in the dynamic nature of criteria (Chielli, 1956). This conventional, yet perhaps erroneous (G. V. Barrett, Caldwell, & Alexander, 1985), belief that performance determinants
change significantly over time logically mandates a greater expenditure of resources for criterion development than many are willing to accept. One result is that little emphasis is likely to be placed on research as a means of accumulating knowledge about the appraisal system (Smith, 1976). For these and many other economic and logistic reasons, it can be assumed that criterion measures are not generally likely to command the concern for rigor necessary for optimal development. Also note that this issue of rigor in behavioral criterion development is not unique to personnel selection research (O'Grady, 1982). Rushon, Brainerd, & Pressley (1983) analyzed behavioral criterion deficiency in 12 major areas of psychological research and concluded that it is a formidable and pervasive problem.

There is a more fundamental conceptual basis for assuming that criterion development is likely to be deficient. As Figure 3 illustrates, there are three inferences linking the psychological construct required for job performance and the operational criterion measure. The truth of Inference 8 is to some extent empirically testable by the construct-related validation procedures discussed earlier. However, Inference 7, linking the performance construct with the underlying psychological construct, is justifiable only through rational deductive analysis (Cascio, 1987). This inference must be based on the judgments of certain people. On the one hand, the criterion must be defined by organization leaders who are responsible for formulating and translating valued organizational outcomes. On the other hand, selection specialists are required to infer from job analytic data the predictor constructs required for job performance. Incidentally, Smith (1976) pointed out that the translation of goals to valued behavior should also be validated (Inference 11). This mandate for collaborative decision making between various professional groups has obvious implications for the quality of the resulting criterion measurement system.

When considering construct-related validation of the criterion, unique conceptual and practical issues do arise. On the predictor side, a test is constructed to sample certain criterial behaviors (Messick, 1980) that are specified by the psychological construct theory and judged to be indicators of a specific construct or set of constructs. Criterion measures, likewise, are developed to be samples of an underlying behavior domain. Notice that the relative position of the psychological construct domain and performance domain has been changed in Figure 3. The framework proposed here portrays psychological constructs as being more deeply embedded in the nomological network and are more fruitfully conceptualized as labels for behavioral regularities that underlie behavior both sampled by the predictor and in the performance domain as sampled by the criterion.

Delineating the Performance Domain

Two prevalent ways of conceptualizing performance domains are discussed in the performance appraisal literature. One school of thought places relative emphasis on a conceptualization of performance domains as collections of overt job behaviors (e.g., Borman, 1983), whereas the other places relative emphasis on outcomes or results (e.g., Kane, 1986). The former is motivated by concern for developing psychological theories that capture behavioral regularities important to organizational functioning. The latter recognizes the importance of goal attainment to organizational functioning.

We join others in stressing the inextricable relationship between job behaviors and outcomes. We propose that performance domains are composed of behavior-outcome units. Outcomes are valued by the organization, and behaviors are the means to these valued ends. As a result, behaviors take on different value, depending on the value of their consequent outcomes. Therefore, optimal description of the performance domain for a given job requires careful and complete delineation of valued outcomes and the accompanying requisite behaviors (Fine, 1986; James, 1973).

The behavior versus outcome distinction is reflected in the distinction between composite and multiple criterion models. The relative merits of these models have been examined in detail elsewhere (Brogden & Taylor, 1950; J. P. Campbell, Dunnette, Lawler, & Weick, 1970; Carroll & Schneier, 1982; Dunnette, 1963; Guion, 1965; James, 1973; Schmidt & Kaplan, 1971; Smith, 1976; Thorndike, 1949). The important difference between these models is often viewed as whether different types of operational criterion information should be combined or not. For this analysis, however, a more fundamental difference between the two models is the way in which performance domains are conceptualized. The composite criterion model implies a unitary (and often economic) conception reflecting an employee's total worth to an organization. As a result, operational criteria are designed to reflect the underlying domain by sampling the "economic" end products of job behaviors" (Schmidt & Kaplan, 1971, p. 424; parentheses added). In contrast, the multiple criterion model conceptualizes performance as a behavioral domain within which some behaviors are more valuable than others for achieving organizational goals. Operational criteria developed to tap this domain are more behaviorally oriented, focusing on individual incidents or dimensions of actual job behaviors that lead to the attainment of valued organizational outcomes.

It is important to emphasize that these notions of total economic worth versus performance behavior domain are no less hypothetical constructions than, for instance, intelligence or reading ability (Schwab, 1980). The fact that job performance can be described in both ways is reflected in the job analysis literature by reference to job-oriented (what is accomplished) versus worker-oriented (what is done to accomplish) bases for job description (Cummings & Schwab, 1973; McCormick, 1976). However, the relevance of this distinction for criterion development has generally been unsystematically examined. In the discussion that follows, we discuss in greater detail the procedures used to justify Inferences 7, 8, 10, and 11.

Generating Evidence of Criterion Validity

Job analysis provides the evidential basis for justifying Inferences 7, 8, 10, and 11. Most personnel professionals are quick to agree that systematic job analysis provides the prerequisite data base for all subsequent selection activities. Yet, perhaps no other professional activity is better characterized by the idiosyncratic use of unstandardized procedures and lack of general
principles to guide data collection (Tenopyr, 1986). Clearly, the proliferation of job analysis procedures is ample testimony to the conclusion that very little in the way of standard job analytic procedures exists. Regardless of the reasons for this lack of standard practice, one result is a relative dearth of both conceptual and empirical guidelines for adequately justifying the critical inferential linkages (critical Inferences 7, 8, 10, and 11).

Job analysis involves examining job demands and translating them into behavior-outcome units that define the performance domain, and that subsequently match optimal person-job matches possible. Inference 10 represents the extent to which the actual job demands have been adequately analyzed, resulting in a valid description of the performance domain. The process of substantiating Inference 10 is commonly referred to as job description. There are at least two fundamental reasons to suspect the validity of Inference 10 in most applied selection situations. First, fully adequate taxonomies of job characteristics, which are required for proper delineation of the performance domain, have yet to be developed (Fleishman, 1975; Fleishman & Quaintance, 1984). Second, most jobs are accurately characterized as collections of demands with associated behavioral universes with fuzzy, if not indeterminant, boundaries (Weick, 1979), making their unequivocal delineation logically impossible.

Inference 11 represents the extent to which behavior-outcome links have been substantiated. Again, job analysis is the process of discovering and specifying these links. Some job analysis procedures more systematically delineate behavior-outcome links than do others. For example, the critical incidents technique (Flanagan, 1954) formally elicits organizationally valued outcomes and systematically ties job behaviors to these. Functional job analysis also formally assesses these linkages through group interviews of subject matter experts (Fine, 1986). Regardless of which method is used, to the extent that job analysis is conducted without explicating behavior-outcome links, the validity of Inference 11 is suspect.

Inference 8 links an operational criterion with the performance domain. As such, it represents the inference that the operational criterion validly measures the performance domain. This process is commonly referred to as criterion development. When the multiple criterion model is used to guide criterion development, job analysis data in the form of worker-oriented (what is done to accomplish) behavior requirements are useful for justifying this inferential linkage. When the composite criterion model is used to guide criterion development, job analysis data in the form of job-oriented (what is accomplished) behavior requirements are most useful. In either case, justification of Inference 8 typically takes the form of (a) claims on the part of the job analyst that all major behavioral dimensions or outcomes have been identified and are represented in the operational criterion measure (e.g., performance rating instrument or objective index) and, occasionally, (b) psychometric evidence of accuracy or lack of bias in indexes of job performance (Dickinson, 1987; Kleiman & Durham, 1981). In other words, criterion measures are usually validated (i.e., evidence for Inference 8 is generated) by rational, albeit tacit, claims about the content-related evidence of validity. The position advanced in this article is that sole reliance on content-related evidence of criterion validity necessarily means that the evidential base is deficient relative to the numerous other forms of evidence available. This is particularly evident in light of Feldman's (1986) call for a taxonomy of appraisal tasks. He is pointing out that different types of tasks influence the manner in which appraisal judgments are made. He goes on to examine the differences in how these judgments are validated.

The conventional practice of relying solely on single criterion measures and methods, whose content is often rather unsystematically determined, inevitably leads to many validation efforts with questionable criterion validity (Guion, 1976). This issue has been addressed over the years and labeled the criterion problem. However, except for James's (1973) exposition, little in-depth analysis of the conceptual issues surrounding the criterion problem has been advanced. Suffice it to say that in many, if not most, validation situations, the validity of Inference 8 is suspect, which in turn weakens conclusions about the validity of other inferences in the system.

Inference 7 is likewise typically justified by the job analyst's claim that from a specific job analysis, he or she has inferred the requisite psychological constructs that underlie performance. This process is commonly referred to as deriving job specifications. Job analysis data in the form of ability requirements (Dunnette, 1976; Fleishman, 1982; Pearlman, 1980) are useful for justifying this inference. Some important theoretical strides have been made in establishing both theoretical and empirical linkages between job behaviors and underlying attributes (e.g., Fine, 1986; Fine & Wiley, 1971; Fleishman, 1978, 1982; Lopez, Kesselman, & Lopez, 1981; McCormick, 1976). However, in practice, it is not uncommon for Inference 7 to be informally justified by job analysts' judgments. Sole reliance on this inductive approach (Bass & G. V. Barrett, 1981) means that the validity of Inference 7 is suspect whether these judgments are not based on current knowledge of construct-behavior relations and sound reasoning about criterion development. Clearly, Dunnette's (1976) call to link "the two worlds of behavioral taxonomies" (p. 514) is still operative.

In addition, personnel specialists must adopt a broader view of what qualifies as relevant empirical evidence for criterion linkages. For example, a program designed to train critical job skills, which is then evaluated by using job performance criteria, provides criterion-related evidence for Inference 7. The point here is that a wealth of empirical evidence supporting criterion inferences might be more systematically derived from the extant training literature. This is particularly relevant in those cases in which training has altered psychological attributes generally regarded as enduring and less amenable to change. These cases would be more relevant because selection programs are more often designed to assess relatively stable constructs because of the perceived impracticality of trying to change them through training.

Given that Inferences 7 and 8 may often be justified on tenuous evidential bases, the model presented in Figure 3 leads logically to an intriguing conclusion regarding the relative superiority of criterion-related versus construct-related evidence of validity for selection decisions. From this perspective, Inference 5 is a surrogate for the fundamental Inference 9, which links predictor information to an applicant's true performance in the
organization. It is this inference for which validity evidence is ultimately sought. Taking this logic one step further, to the extent that Inference 8 is questionable, empirical evidence of Inference 5 is not as relevant to Inference 9. Sound evidence of Inferences 6 and 7 would provide a much more substantive justification of Inference 9, in this instance. Most selection specialists would find it rather easy to recall situations in which their confidence in Inference 5 (as an index of Inference 9) was severely weakened by information about the deficiency or contamination of a poorly developed criterion measure. Yet, in the same situation, use of an established assessment instrument, in combination with a rigorous rationale for why performance requires certain psychological constructs, would provide a firmer evidential basis on which to conclude the validity of resulting decisions. This is a case in which construct evidence of validity is superior to criterion-related evidence. Although this has been suggested by others, we contend that the lack of critical analysis of Inferences 5 and 8, which characterizes most validation research, has caused a dramatic underestimate of the frequency with which construct-related evidence is judged superior to criterion-related evidence.

In the next section, the personnel selection framework presented in Figure 3 is broadened. Recommendations for future conceptions of the validation process are then discussed in this context. It is important to remember that although terminology will change somewhat and different inferences will be emphasized, the process is essentially one of traditional construct validation. It should become clear from this perspective that the science of psychology as applied to personnel decision making involves the development of theories, validation of constructs, and generation of evidence to support important inferences about people and their behavior at work.

The Psychological Science of Personnel Decision Making

Our contention is that the validation process discussed thus far, if adequately adapted to the unique needs of personnel selection, provides a broader framework for expanding conceptions of validity. Thus far, however, our discussion of critical inferential linkages has resulted in a focus on a more narrow conception of the validation process than is ultimately desirable. We now present a broader view of the nomological framework relevant for developing theory within personnel psychology. This framework is schematically presented in Figure 4. The left side of Figure 4 represents the more traditional notion of construct validity described by Cronbach and Meehl (1955). The center of Figure 4 represents the focus of greatest interest to applied decision makers. Inferences 5, 6, 7, and 8 are of particular concern because of their direct relevance for justifying Inference 9. That is, they are relevant for determining the extent to which inferences from scores on a test of some predictor construct allow predictions of actual job behavior.

Inference 9 is of utmost importance to applied decision makers. Empirical evidence of Inference 5 provides partial support of Inference 9 and can be conceived of as a special case of Inference 13. Evidence supporting Inference 5 can be direct and take the form of empirically observed relationships. Messick (1980), when discussing construct-related validity, stated that "some of the construct's nomological relations thus become criterial when made specific to the applied setting" (p. 1019). He added that these predictive relationships are singled out for special attention under the rubric of criterion-related validity and differ from general nomological relations in being more narrowly focused on specific sets of data and specific applied settings. Similarly, Cook and D. T. Campbell (1979) explicitly stated that priorities regarding validity issues are fundamentally different between theoretical and applied researchers. From this perspective, it can be seen that criterion-related validity evidence is appropriately viewed as a special type of convergent (or discriminant) evidence of construct-related validity. One can also generate various indexes of content overlap to support Inference 9. Evidence of Inference 9 can also be indirect and take the form of convergent and discriminant relationships between the components linked by Inferences 6, 7, and 8. Extending this logic beyond the original focus, Inference 6 is strengthened by evidence of Inferences 11, 12, 13, and so forth.

The right side of Figure 4 represents construct-related validity of the criterion. As mentioned earlier, performance domains have traditionally not been as deeply embedded in networks of theoretical relationships as constructs on the predictor side. Theorists and practitioners need to be increasingly aware of the need to empirically investigate linkages of Inferences 5, 16, 17, and 18 as evidence to support Inferences 7, 8, 14, and 15 (Vance, MacCallum, Coover, & Hedge, 1988).

Traditionally, the focus has been almost exclusively on Inference 5 through correlations between test scores and a criterion measure, occasionally on Inference 17 through development of alternative criterion measures (e.g., Alexander & Wilkins, 1982; Cascio & Valenzi, 1978; Holzbach, 1978; Lee, Malone, & Greco, 1981), and on Inference 8 through assessment of rating accuracy by "true scores" of performance domains (e.g., Bernardin & Buckley, 1981; Borman, 1979; Hedge & Kavanagh, 1988). Recently, Heneman (1986), in a comparison of supervisory ratings and results-oriented performance indexes, called for greater emphasis on convergent evidence of criterion validity. Similarly, James (1973) called for more emphasis on the three levels of criterion measurement proposed by J. P. Campbell et al. (1970), namely, job behaviors, results, and organizational outcomes (Smith, 1976).

A Social-Cognitive Perspective on Job Analysis and Criterion Development

We join others in calling for renewed interest in more rigorous, conceptually coherent criterion development. One important issue is that job analysis efforts need to be directed more at capturing the reality of the organizational context in which criterion judgments are actually made (Feldman, 1986; Stern, Stein, & Bloom, 1956; Wiggins, 1973). Our contention is that typical job analyses produce information that is useful for developing explicit performance criteria, yet is potentially irrelevant to the implicit criteria that often may be used to evaluate day-to-day performance or promotability (Turnage & Muchinsky, 1984). To the extent that the validity of Inference 10 is questionable, all other inferences in the system are questionable.
Increased concern for the validity of Inference 10 has motivated considerable research in recent years on a variety of social and cognitive factors affecting job analysis data (Arvey, Davis, McGowen, & Dipboye, 1982; Cornelius, DeNisi, & Blencoe, 1984; DeNisi, Cornelius, & Blencoe, 1987; Friedman & Harvey, 1986; Green & Stutzman, 1986; J. E. Smith & Hakel, 1979). The validity of both Inferences 7 and 8 is dependent on job analytic data used to translate performance behavior into measurable criterion elements and to delineate overlap with predictor constructs. Better understanding the judgments that underlie perceptions of jobs and performance may contribute to improved criterion development (Guion, 1986). In other words, the basic cognitive processes that underlie perceptions of people may also underlie the perceptions of jobs, and therefore the vast research on person perceptions can be integrated and generalized to enhance our understanding of the determinants of job perceptions (e.g., Binning, Zaba, & Whattam, 1986; Cantor & Mischel, 1979; Cleveland & Landy, 1983; Cooper, 1981; Feldman, 1981, 1986; Funder, 1987; Lord, 1985a, 1985b; Swann, 1984).

A particularly integrative approach is exemplified by Cantor, Mischel, and Schwartz's (1982) prototype assessment of psychological situations. They assessed peoples' prototypical beliefs about person–action combinations in common situations and found considerable consistency across people. This approach could be adapted to the study of job prototypes and their effects on job analysts' perceptions of as well as incumbent performance. Similarly, a considerable amount of job design research exists showing that social information affects perceptions of task characteristics (Salancik & Pfeffer, 1978; Thomas & Griffin, 1983; Weiss & Shaw, 1979).

At a more molar level, criterion development is largely a sociopolitical process and therefore deserves greater attention from this perspective (e.g., Katz & Kahn, 1966; Longenecker, Sims, & Gioia, 1987; Mitchell & Linden, 1982; Weick, 1979). Programs designed to train raters of performance could also benefit from the integration of research described earlier (Bardes & Buckley, 1981; Borman, 1979; Landy & Farr, 1980).

Expanding on the framework presented, Inference 16 might involve relating specific performance criteria to measures of nonjob behaviors that are theoretically expected to relate to performance behaviors in some specified way (e.g., Blau, 1985; Rousseau, 1978; Youngblood, 1984). Youngblood's (1984) study of work and nonwork explanations for absenteeism might exemplify this approach. He found that absenteeism could be explained by the importance of leisure activities engaged in away from the work setting. In a similar vein, perhaps successful managers are more proficient at organizing successful family vacations than are their less competent organizational counterparts. The point we are making here is that the delineation of behavioral domains can be conceptualized by reference to the theory surrounding psychological constructs (predictor side) or "theory" surrounding job performance (criterion side). Meaningful behavioral regularities may be discovered by investigating relationships between either type of behavioral domain.

Methods used by personality and social psychologists could be adapted for the study of these criterion relationships. For instance, Mischel (1984), Funder (1987), and Funder and Colvin (1988) reviewed studies relating both lay perceptions and objective measures of personality to independent measures of behavior gathered from peers and family members. Relationships between work and nonwork behavior could be investigated in an analogous manner. Similarly, the logic of biographical data could be adapted to criterion research. Although biodata instruments are typically used for prediction purposes, data about nonwork behavior could be collected using an analogous questionnaire format. For example, in attempting to verify the accuracy of biographical information, Shaffer, Saunders, &
Another Call for Experimenting Organizations

Perhaps the greatest advancement for the science of personnel psychology will come only when the values driving organizational administrators’ decisions about behavioral science research are changed. For many reasons, the behavioral sciences have what Staw (1977) described as a “center–periphery” relationship to the administrative users of scientific knowledge. He contends that new knowledge is created by researchers “who are presumably at the center of knowledge” (p. 426) and it is considered their responsibility to disseminate expertise to organizational users in a prepackaged, “formula-like” fashion. Failures of behavioral science interventions are thus more likely to be attributed to deficiencies in science and knowledge rather than inappropriate expectations for generalizability. Staw (1977) envisioned a much healthier relationship in which the seat of innovation is at the local organizational level. This shift in values reorients the role of behavioral science research so that it lies in the periphery as a resource to guide organizational experimentation. Concomitant with this reorientation is a shift in the educational role of behavioral science and the manner in which knowledge is disseminated. Rather than persuading practitioners to adopt a particular theory or planned intervention, psychologists’ efforts should be directed more toward selling the benefits of experimenting organizations, where ongoing, systemwide, multivariate research is made an integral part of organizational functioning. Consequently, the psychologist’s role would increasingly involve training practitioners in research evaluation skills.

A concomitant shift from summative to formative evaluation (Staw, 1977) is also desirable. The process of inferring whether a specific program or intervention has worked or has not had a positive effect is referred to as summative evaluation. A more iterative and ongoing process of selecting program goals and building organizational interventions is referred to as formative evaluation. Formative evaluation implies the successive approximation of desired organizational systems, built through a series of trials in which failures are considered as informative as successes. In this new research context, the term failure merely implies some unpredicted outcome or result, equally useful for refinement in the next stage of program development. It is necessary to change typical organizational values so that systems can be developed to effectively monitor, provide feedback, and utilize negative as well as positive data.

The creation of experimenting organizations could have vast implications for personnel selection research. Greater emphasis would be placed on large-scale, programmatic research involving the melding of laboratory and field settings (G. V. Barrett, 1972; Flanagan & Dipboye, 1981). This could lead not only to richer, more efficient theory development but also better understanding of longitudinal changes in employee and job characteristics. For instance, Helmreich, Sawin, and Carsrud (1986) demonstrated the effects of predictor–criterion time lags on the predictive power of personality characteristics. They argued that personality traits have their most potent effects on job performance only after considerable time on a job. Longitudinal data gathering is important for criterion measurement as well. Meyer (1987) presented data suggesting that cognitive ability tests are more predictive of managerial promotional progress over time than of supervisory ratings of performance at a given point in time. The same general issue can be raised regarding social influence effects that result from any organizational intervention. Administrators’ preoccupation with one-shot, short-term identification of successful selection procedures has most likely masked many useful approaches to employee selection.

Experimenting organizations also would create an environment in which macro and micro issues could be more systematically integrated. For example, little work has been done to create contingency models relating organizational structures with job design and criterion development processes. Yet, Mintzberg (1983) describes in detail how organizations’ structural parameters affect individual-level control mechanisms and job characteristics. Some organizational structures (e.g., machine bureaucracies) contain jobs that are more amenable to the multiple criterion model. The primary coordination mechanism in machine bureaucracies is standardization of work processes. This is possible because jobs are highly routine, and behavior–outcome links are explicitly understood and programmed. In other types of organizations (e.g., professional bureaucracies, ad hoc organizations) the composite criterion model may be more appropriate. In these types of organizations (or parts of organizations) where work processes are more complex and less programmable, coordination is achieved through other mechanisms such as standardization of work output, necessitating a performance domain that comprises outcomes rather than behaviors. In still other types of jobs such as higher level managerial jobs, neither work processes nor work outputs are specifiable a priori (J. P. Campbell, 1983; Feldman, 1986; Palermo, 1983). In these instances, coordination is achieved primarily through standardization of input skills and knowledge. In these jobs, an individual’s worth to the organization might be more fruitfully indexed by assessment of changes in job-related knowledge. This logic underlies trait-based and competency-based assessment (Sokol & Oresick, 1986). We note that it is conceptually flawed to refer to trait-based performance appraisal, because job performance is not being appraised. Personal characteristics are being assessed with the assumption that they will reflect an individual’s worth to the organization. Rather than assessing an individual’s
contribution to organizational goal attainment, the potential of such attainment is assessed by using the same logic as personnel selection.

Micro–macro integration is desirable on the predictor side as well. A basic tenet of Schneider's (1987) attraction–selection–attrition framework is that macro–organizational structure differences can be best understood through types or profiles of individual employee characteristics. We believe that the creation of experimenting organizations will ultimately do more to enhance our movement from test validation to selection research (Guion, 1976), making it more likely that dynamic, multivariate relationships can be fruitfully understood and used to enhance the quality of staffing decisions and ultimate organizational effectiveness.

References


Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations:


Received December 8, 1987
Revision received September 23, 1988
Accepted October 4, 1988