

CHAPTER 1

Development of Validation Evidence

S. Morton McPhail, Valtera Corporation

The owner of a small restaurant chain with three locations employs five food servers at each location. As repeat business in the neighborhoods near her restaurants is vital to her success, she is concerned about the level of customer service being provided by her current employees. She has two openings to fill and wants to do a better job of selecting people who will display the cordial, attentive, helpful behaviors she believes are so crucial to her business' success.

At a large petrochemical processing facility, many of the maintenance mechanics were hired at about the same time and will reach retirement age in the next year or so. Although the plant is quite large, the company has been contracting increasing amounts of the maintenance work. Currently, there are thirty maintenance mechanics, thirty plant electricians, five machinists, and twenty-five pipefitters employed at the facility. The plant superintendent recognizes that when the company hires replacements for the mechanics, it will be hiring at the journey- level, but he wants to make sure that the new employees will be able to learn the plant's equipment and become effective quickly. His own experience suggests that mechanical aptitude will be an important selection requirement. He has told the human resources manager to implement use of such a measure for the upcoming round of hiring.

A large manufacturer of food products anticipates a need to hire plant operators in the near future. The company has a dozen

plants scattered across the United States, but there are fewer than thirty operators at any single facility. The company is seeking to migrate to team-based operations, and Human Resources has determined that strong interpersonal skills have been important factors in the success of a pilot of the team-based model.

An aluminum manufacturer is preparing to hire workers at a new, state-of-the-art rolling mill. The “greenfield” plant incorporates new technology and work techniques that are unlike those used at any other such facility in the world. The company will need to hire some three hundred workers into multicraft, multiskilled positions before the plant is operational to ensure that training can be completed prior to initial startup. The human resource director anticipates that because of the desirability of the working conditions and the reputation of the company, there will be over six thousand applicants.

A public safety agency has a long and troubled history of dealing with representatives of the employees’ bargaining unit. There is a significant level of mistrust among the employees concerning management’s intentions. Anticipating a need to select and hire new employees from among a very large applicant pool, the agency (which has faced substantial litigation in the past) wants to be sure that the selection measures comply with legal requirements. However, despite sincere reassurances, employees are reluctant to participate in experimental testing, and supervisors are hesitant to make meaningful evaluations of subordinates based on the defined performance criteria.

Many practitioners will recognize these or similar situations. These scenarios include large and small organizations in both manufacturing and service and in both the public and private sectors. They include concerns about both cognitive and noncognitive individual differences. They represent many of the issues that make the selection problems faced by organizations distinct and unique, yet in some ways similar. In what way, then, are these scenarios similar? Is there a common theme here? Despite their apparent differences, the underlying dilemma these situations present for selection research is the same. Given the constraints inherent in practical applications and field research, how are we to develop evidence supporting the interpretations of the selection measures we use in these and many other diverse, yet similar situations? This volume offers some answers to this dilemma.

As the epistemological character of what validity means in the selection context has matured, the nature of the research necessary to evaluate it as a basis for the use of, and reliance on, selection procedures has become both more varied and conceptually complex. This complexity both raises constraints and offers opportunities. On the one hand, we have come to understand that straightforward correlational approaches sometimes are simply not available to us and in many instances raise questions of scientific feasibility and veracity. On the other hand, a broadened understanding of validity as evidence-based hypothesis testing offers an array of strategies for seeking and obtaining relevant evidence. Before exploring issues (some of which are illustrated by the opening scenarios) that constrain and complicate validation research, I begin by examining the concept of validity and validation itself. Subsequent chapters examine alternatives for obtaining and evaluating evidence bearing on the question of validity.

The Evolving Conception of Validation

Early selection research focused on what has come to be called *dust bowl empiricism*. Far from addressing the underlying questions of why a particular measure or assessment provides meaningful information about future job behaviors, most researchers were satisfied to demonstrate that a functional relationship existed between the test measure and performance indicators (Schmitt & Landy, 1993); that is, the empirical finding was considered sufficient to sustain the inferences made on the basis of the test scores.

From 1950 to 1954 the American Psychological Association (APA) undertook to codify requirements for technical justification and publication of tests, culminating in the Technical Recommendations for Psychological Tests and Diagnostic Techniques (APA, 1954). These recommendations became the basis for the *Standards for Educational and Psychological Testing*, now in its fifth version (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). In its recommendations, the APA Committee on Psychological Tests in 1954 identified four categories of validation evidence: predictive, concurrent (which became lumped with predictive as criterion validation), content, and construct (a category subsequently explicated by Cronbach & Meehl, 1955). These

categories came to be widely discussed as “types” of validity, and some effort was made to define the situations and purposes for which each provided meaningful evidence of validity (see Landy, 1986).

Initially, there was some confusion about what was meant by the term *construct validity*. However, seminal articles by Cronbach and Meehl (1955) and Campbell and Fiske (1959) were instrumental in clarifying the meaning of the concept and providing research methodologies to evaluate the extent to which a test demonstrated this “type” of validity. Key concepts such as the nomological network (the consistent web of theoretical and empirical relationships into which a construct systematically fits) (Cronbach & Meehl, 1955) and convergent and discriminant validity and multitrait, multi-method matrices (Campbell & Fiske, 1959) were introduced and ultimately adopted widely in the field of psychology.

The tripartite validity typology provided a convenient way for psychologists to discuss how one would go about the scientific process of providing evidence to support a proposed interpretation of a test score. But the “law of unintended consequences” was lurking in the unexplored woods of civil rights litigation. With the passage of the Civil Rights Act of 1964 and the creation of the Equal Employment Opportunity Commission (EEOC), tests would be subjected to new and intense scrutiny as being potential barriers to equal opportunities for groups now protected under law from discrimination. The EEOC’s original *Guidelines on Employee Testing Procedures* (1966) and *Guidelines on Employee Selection Procedures* (1970) borrowed from the *Standards* and for the first time incorporated technical testing standards into governmental regulations. By 1978, with the adoption of the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978) by the federal equal employment enforcement agencies, the “trinitarian” (Guion, 1980) conception of validity was firmly enshrined in both regulatory rules and case law.

Thinking about the meaning of validity and validation continued to evolve. It had never been the intent of those who conceived of validation as falling into three categories to limit the concept to these types. Guion (1980), following Loevinger (1957), pointed out that this conception was unacceptably limiting. Landy (1986)

as well as others (such as Cronbach, 1971, and Messick, 1988) reframed the issues by firmly placing validation as a scientific process in the mainstream of hypothesis testing. Validation, Landy argued, is directed toward justifying the inferences drawn about people based on a test score. Thus, to the extent that the inference we are seeking to make is that job performance is a function of the knowledge, skills, abilities, and other characteristics (KSAOs) reflected in a test score, we are engaged in the process of accumulating evidence to test the hypothesis of the existence of a functional relationship against the null of no relationship. Thus, even if the goal is to draw inferences about future job performance, “it does not necessarily follow that criterion-related validation strategies are the only means for documenting the soundness of those inferences” (p. 1187). Landy concluded that psychologists must reject the artificial distinction made in the *Uniform Guidelines* between behavior and mental process and instead begin thinking of validation research as hypothesis testing. As scientists, psychologists are trained to conduct research to test just such hypotheses.

In a retrospective article, Guion (1987) made note of a number of trends that he had observed in validation research. He reviewed changes in predictors, criteria, data collection methods, and validity and validation. In discussing changes in the conception of validation, he made a very pragmatic distinction between job-relatedness and construct validity, noting that “[a] variable that reliably predicts a job-related criterion is job related, even when one does not know what it measures. Validity of measurement is a psychometric question; it requires, in contrast, a clear idea of the construct being measured . . .” (Guion, 1987, p. 212). This distinction is of particular importance for practitioners who often lack the resources and opportunity to apply the rigorous, programmatic research implied by Messick’s (1988, 1989) more scientifically orthodox discussions of validity.

In a symposium paper subsequently published in Wainer and Braun (1988), Messick offered what may be the most quoted definition of validity in the unified perspective (author’s emphasis):

Validity is an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores. (p. 33)

He argued that three issues were foundational to the concept of validity: (a) “interpretability, relevance, and utility of scores,” (b) “import or value implications of scores for action,” and (c) “functional worth of scores in terms of social consequences of their use” (p. 33). As an aside, it should be noted that this final point did not go unchallenged, and the concept of “consequential validity” has met with considerable debate, with cogent arguments being made both supporting and attacking it (see Shepard, 1993, 1997; Lees-Haley, 1996; Zimiles, 1996; Linn, 1997; Mehrens, 1997; and Popham, 1997).

Messick (1989) decried what he described as the “disjunction between validity conception and validation practice” (p. 34). He argued that “construct-related evidence undergirds not only construct-based inferences but content- and criterion-based inferences as well” (p. 40), setting a standard for validation research that seems to require a programmatic approach to validation. Like Landy (1986), Messick (1988) viewed validation as a fundamentally scientific endeavor:

Test validation in the construct framework is integrated with hypothesis testing and with all of the philosophical and empirical means by which scientific theories are evaluated. Thus, construct validation embraces all of the statistical, experimental, rational, and rhetorical methods of marshaling evidence to support the inference that observed consistency in test performance has circumscribed meaning. (p. 41)

In a similar vein at the same symposium, Angoff (1988) made it clear that “. . . construct validity as conceived by Cronbach and Meehl cannot be expressed as a single coefficient. Construct validation is a process, not a procedure; and it requires many lines of evidence, not all of them quantitative” (p. 26).

This viewpoint had, of course, always been at the heart of the notion of construct validity but was obscured by the more easily understood tripartite typology, which in applied practice led to the notion of the need to make a choice between types of validity or appropriateness of a particular strategy. But as the unitary conception tied to the scientific process was being reasserted more and more strongly, the fly in the ointment continued to be the codification of the threesome into the *Uniform Guidelines*, a fact that was not going to change any time soon.

In a subsequent chapter in Linn (1989) (and in a later paper in *American Psychologist* [Messick, 1995]), Messick expanded on this position and furthered his theme of the importance of taking into account the social impacts of testing in considering the validity of tests. He asserted that the justification for testing could be divided into two facets: evidential and consequential. The consequential facet (or basis)

. . . of test interpretation is the appraisal of the value implications of the construct label, of the theory underlying test interpretations, and of the ideologies in which the theory is embedded. . . . [V]alue implications are not ancillary but, rather, integral to score meaning. (p. 20)

In that seminal chapter, Messick (1989) traced the philosophical history of validity contrasting and integrating the “verifiability” requirements of logical positivism (Ayer, 1935/1946/1952) and the “falsification” principle of Popper (1935/1959/2002). In a subsequent, essentially philosophical, analysis, Markus (1998) described this integration (as had Messick) in terms of a Hegelian synthesis. In making these arguments, Messick noted that the early development of concepts of construct validity owes much to the positivist position. The notion of the development of a verifiable nomological network as a basis for claims of construct validity arose from this same tradition. He cited Cronbach (1989) (in press at the time) as stating: “it was pretentious to dress up our immature science in the positivist language; and it was self-defeating to say . . . that a construct not a part of a nomological network is not scientifically admissible” (Messick, 1989, p. 23).

With Messick’s (1989) chapter, the unitary conception of validity was firmly established. In that same year, Schmitt (1989) similarly defined construct validity as “the degree to which certain explanatory psychological concepts or constructs account for performance on a test” (p. 332). He argued for greater attention to such validity in both tests and performance indices. He then proceeded to review literature demonstrating a variety of means by which evidence relevant to the validity of a purported measure of a construct could be developed.

Binning and Barrett (1989) reinforced the unitary conceptualization of validity. They expanded Nunnally’s (1978) model of critical linkages in validation by articulating and distinguishing

relationships among the elements of the logical system composed of the inferences that underlie the unitary concept of validity. They extended the concept of construct validity to the performance domain and emphasized the role of job analysis as an integral part of the validation process. Concluding with a call for “experimenting organizations,” they recommended a process of “formative evaluation,” which they defined as implying “the successive approximation of desired organizational systems, built through a series of trials in which failures are considered as informative as successes” (p. 490).

Kane (1992) presented the case for an “argument-based approach to validity” in which the plausibility of a test score interpretation depends on the existence of evidence supporting “various lines of argument in specific contexts, the plausibility of assumptions, and the impact of weak arguments on the overall plausibility” (p. 528) of the proposed interpretation. Such arguments may be theory-driven, but it is not necessary that they be so. This approach extended the “hypothesis testing” paradigm to better identify and specify what hypotheses (regarding the argument itself, its cohesiveness, and its underlying assumptions) need to be tested to support the inferences required by the interpretation.

Arvey (1992) sought to clarify conceptual and research issues involving construct validation. He defined construct validity as “providing, acquiring, developing, or otherwise establishing information or data to decision makers that an operational measure does indeed reflect the construct that is thought to underlie the measure” (p. 61). With this definition, he conceptualized construct validation as a special case of model building and testing in which “there is a reliance on both empirical data and analytic methods and rational deductive logic to confirm the model” (p. 65). One might note here some similarities to the argument-based approach proposed by Kane (1992). Arvey proposed a number of practical research methods that might be thought of as construct validation procedures for how one could go about developing evidence for test validity. Landon and Arvey (2006) revisit and extend this discussion in this volume.

Contemporaneously, Geisinger (1992) traced the history of validation thinking and identified ten ways that the concept of validity has changed (loosely paraphrased):

1. From emphasis on the test *to* emphasis on evidence to support specific inferences from the test
2. From atheoretical *to* primarily theory-based
3. From grand nomological networks *to* “micromodels”
4. From behavioral criteria *to* limited scope theories
5. From focus on the test developer as arbiter of validity *to* users and decision makers as responsible for the validity of test use
6. From the tripartite view—with construct validation as equal to (or even slightly less than) criterion-related validation—to a view of construct-related validation as incorporating other evidential bases for evaluating validity
7. From the (perhaps still ongoing) debate about content validity as a form of validation *to* content relevance and content coverage as important evidential bases for construct validation
8. From concurrent and predictive validity *to* criterion-related validation as an aid to applied decision making instead of formal validation
9. From situation specificity *to* validity generalization
10. From single validation analyses *to* meta-analyses and structural equation modeling

In a volume of the *Frontiers* series, Schmitt and Landy (1993) recapped the development of the concept of validation as testing the inferences required by the Nunnally (1978) model, which had been further articulated by Binning & Barrett (1989), arguing for all validation research to be theory/construct grounded. They examined “construct-relevant research” as it occurs in the selection context, pointing to the centrality of job analysis (as had Binning and Barrett, 1989) as part of the evidential basis for validation. They further considered research requirements for both criteria (long overlooked) and predictor constructs (of various kinds) with emphasis on the theoretical linkages between predictors and the performance domain. They concluded with a discussion of alternate methods of validation—all of which are considered in greater depth in this volume.

Continuing the theme he had broached earlier (Guion, 1987), Guion (1998) differentiated between “psychometric validity,”

referring to “[i]nterpretive inferences [that] describe characteristics revealed by the measurement procedure itself” (p. 237), and “job relatedness,” which are “[r]elational inferences interpret[ing] scores in terms of different but correlated characteristics” (p. 237). He then described four bases of evidence that bear on the “*evaluation* [emphasis in original] of tests and test uses” (p. 239): (1) test development, (2) reliability, (3) patterns of correlates, and (4) outcomes.

From 1993 to 1999, representatives from the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council of Measurement in Education (NCME) served on a joint committee to review and revise the 1985 *Standards for Educational and Psychological Testing*. The resulting *Standards* (American Psychological Association, 1999) continued the precedent set by the 1985 version of referring to “types of validity evidence” rather than to “types of validity.” The language in the discussion of validity evidence continued to reject the tripartite view: “Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose” (p. 11). The section goes on to discuss different sources of validity evidence that resemble Guion’s (1998) outline, including evidence based on (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing.

Over the two-year period from 2000 to 2001, a task force appointed by the president of the Society for Industrial and Organizational Psychology (SIOP) worked to review and revise the third edition of *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 1985). The task force was charged to review and update the 1985 *Principles* to ensure that they continued to reflect the state-of-the-science in I/O psychology and to make them consistent with the then recently issued *Standards*. Their work resulted in the fourth edition of the *Principles* adopted by SIOP and APA in 2003. The 2003 *Principles* continued to endorse the unitary concept of validity, embracing and quoting the definition adopted by the *Standards* regarding accumulation of evidence supporting test interpretations. The *Principles* elaborated on this definition as follows:

Because validation involves the accumulation of evidence to provide a sound scientific basis for the proposed score interpretations,

it is the interpretation of these scores required by the proposed uses that are evaluated, not the selection procedure itself. (p. 4)

The subsequent discussion went on to specifically reject the tripartite view of validity and to adopt the unitary concept as stated in the *Standards*. To this end, the *Principles* explicitly addressed the issue of types of evidential bases for supporting validity:

... even when different strategies are employed for gathering validation evidence, the inference to be supported is that scores on a selection procedure can be used to predict subsequent work behaviors or outcomes. Professional judgment should guide the decisions regarding the sources of evidence that can best support the intended interpretation and use. (p. 5)

Thus, our current conceptualization of the nature and purpose of validation has matured and changed. Validation is seen as embracing both science (“scientific basis”) and practice (“can be used to predict”).

The Scientist-Practitioner Dilemma

With this background, we come to a dilemma that we face today. As scientist-practitioners, we find ourselves serving two purposes having different goals. On the one hand, we are engaged in theory building that involves hypothesis testing, as Landy (1986), Messick (1989), and Binning and Barrett (1989) described. On the other, we are engaged in application and demonstration of job relatedness that often has a different standard of success.

Theory Building

The philosophy of science underwent a number of substantial developments in the twentieth century. A philosophical position associated with the Vienna circle of philosophers in the 1920s and 1930s became known as logical positivism and reached its apogee in the work of Ayer (1935/1946/1952). In the purest form of logical positivism, the positivists argued that propositions had to be verifiable and verified through empirical evidence in order to have status as meaningful (Messick, 1989; Ayer, 1935/1946/1952). This position met with substantial reaction, especially from Karl Popper

(1935/1959/2002), who argued for a principle of falsification (Cohen, 1994/1997; Thornton, 2005) rather than verification.

Following Popper, as Hubley and Zumbo (1996) put it, “The essential logic of construct validation is disconfirmatory. That is, one is trying to show that alternative or competing inferences do not destroy the intended interpretation” (p. 211). It is not possible to assert the verification of a statement of validity with certainty, but it is possible to assert that empirical data contradict it.

The rationale for falsification as the basis for statistical inference lies in Ockham’s Razor (Rindskopf, 1997), which states that “entities are not to be multiplied beyond necessity” (Blackburn, 1996); that is, the most parsimonious explanation for an observation is likely to be the best explanation. Thus, we hypothesize no effect or no relationship (the simplest theory) until that theory is falsified by the reliable observation of an effect different from that simplest theory, and we are thus obliged to accept the existence of a more complicated explanation.

Theory building, then, becomes a process of postulates subjected to “test to failure.” No statement or theory can be fully proven; rather, it must be viewed as not yet rejected as false (Messick, 1989; Popper, 1935/1959/2002). No fault attaches to the assertion of a theoretical position that is falsified; rather, falsification itself is viewed as a step on the path to better, more accurate understanding. Indeed, theories that are not subject to falsification do not have the status of scientific theories at all. Failure to sustain our predicted theoretical outcomes is not failure, but progress to a new and better theory.

Application and Demonstration

In actual practice, however, we are seldom asked by organizations to evaluate data from a falsification viewpoint; that is, we are asked to “validate” selection procedures, not to falsify the null hypothesis. We are asked to demonstrate—or even, in some cases, just to document—the validity of those procedures. Indeed, in the legal arena we are often asked to prove the validity of our tests, and it is cold comfort on cross-examination that the most we can say is that the data are incompatible with a zero relationship. To some extent the process appears at least akin to validating a parking receipt or, as Landy (1986) put it, “stamp collecting.”

The Challenge

Is what we do in practice, then, science? Messick (1988) noted this conflict somewhat obliquely in a comment about applied science: “Moreover, the practical use of measurements for decision making and action is or ought to be *applied* [emphasis in original] science, recognizing that applied science always occurs in a political context” (p. 43).

It would seem that the answer to this question depends on the rigor (both operational and intellectual) that we bring to the endeavor. To the extent that we use the methods of science to reduce the chance that we will be self-deluded into false conclusions based on what we *want* to be the case rather than what *is*, we can lay claim to the rubric of science. However, to the extent that we simply misuse the trappings of science to support our self-serving (or even heartfelt) conclusions, we are charlatans putting on airs.

The following chapters do not address the traditional trinitarian sources of validation evidence. They do not discuss issues of criterion studies or content validation of work samples or simulations. The chapters in this volume deal with the practical realities of validation research. For example, Chapter Nine, “Practical Construct Validation for Personnel Selection,” deals only tangentially with tying test measures into the nomological network; it chiefly addresses concrete methods and techniques for evaluating the inferences implied by the use of a selection measure to predict future performance. In most cases, the evidential methods described in this volume are appropriate for situations like the vignettes at the beginning of this chapter in which traditional (in particular, criterion-related) validation research is not possible or practical; however, they are also evidential bases in their own right. Viewed from Binning and Barrett’s (1989) model of inferential requirements, each of these strategies can be considered part of the larger scientific process of validation: adding bit by bit to our knowledge, testing hypotheses, reforming our models, and testing them yet again—in short, conducting “formative evaluation” (Binning & Barrett, 1989). Some of the strategies offer small steps (such as extending existing knowledge into new spaces), others offer broader information (such as may be gained from assembling large samples in consortia), but all of them can be seen as contributing in some way to the validity evidential base. Moreover, these strategies offer recommendations for

conducting the relevant research in ways that can stand up to professional scrutiny and comply with regulatory requirements.

Practical Constraints

There are a variety of practical constraints that affect the research that I/O psychologists conduct in organizations. Some of these constraints will impact any selection research done in organizations, from job analysis to criterion validation, including some of the alternatives presented in the following chapters. Most, however, are particularly relevant to criterion-related and some construct (for example, experimental) research designs. Many of these constraints lead to effects addressed by Campbell and Stanley (1963) and Cook and Campbell (1979), such as threats to internal or external validity; others, however, may preclude or limit the conduct of the research itself. A number of these constraints are illustrated in the vignettes that opened this chapter.

For ease of presentation, I have organized these constraints into five categories. In the discussions that follow I have described each category and provided examples of the kinds of constraints implied. However, it should be clear that the categories are not necessarily mutually exclusive and, further, that this discussion is by no means an exhaustive compendium of all of the practical constraints faced by researchers in applied settings.

Scientific

A number of the constraining factors are somewhat technical in nature. I have termed them *scientific* constraints because they raise questions about the veracity of conclusions reached from the research that is conducted. At the top of this list is our inability (or perhaps unwillingness) to articulate “strong” hypotheses, thus failing to ask questions of sufficient relevance to the types of inferences of greatest interest. This point has been made by many, but possibly none more cogently and pointedly than Cohen (1994/1997). By failing to specify our hypotheses correctly (both in terms of the contingent probabilities and the comparative effect sizes), our hypothesis testing becomes weak, constraining the conclusions we may legitimately draw from them. For a more in-depth examination of this issue see Harlow, Mulaik, and Steiger (1997).

Perhaps most ubiquitous among the scientific concerns is that of sample availability. Schmidt and Hunter (1998) pointed out that investigations of distributions of correlations have identified small sample size as a major source of error, resulting in what Guion (1987) earlier recognized as “far fewer situations where local validation studies are considered feasible” (p. 206). Other authors (such as Messick, 1989) have suggested that small sample validation studies actually provide less—and less reliable—information than that available from meta-analytic research, a view adopted in part (albeit with substantial caveat) by the *Principles*. Sackett and Arvey (1993) addressed issues of selection in situations that provide only small samples. Some of the methodologies that they discussed are included in this volume, with considerable expansion. Their conclusion about the value of conducting limited research in such settings is worth noting. They sought “to frame the issue as one of incremental improvements over the haphazard selection done in many organizations” (Sackett & Arvey, 1993, p. 445). This same theme is reflected in many of the chapters to follow with a continued emphasis on the need for careful rigor, which is, if anything, greater when seeking “incremental improvements” in the absence of substantial datasets.

Size, however, is only one of the constraints faced when we consider samples for our research. Indeed, especially with respect to concurrent validation research, even the mere existence of appropriate samples is often problematic. For example, long-tenured, highly selected workforces are certainly not ideal for generalizing to the applicant pool expected for future job openings, but those may be the only employees available for data collection. Moreover, it is often the case that the in-place workforce lacks sufficient demographic diversity to allow for meaningful evaluation of subgroup differences or test fairness. Certainly other designs (such as longitudinal studies) could address some of these issues, but as is discussed in the following passages, few employers are prepared to invest the time and resources necessary for such analyses.

A related but different issue involves the predicate for virtually all of our statistical models: random selection and assignment. If scientific research is to be able to attribute causal effects, then our research designs must allow rejection of a null hypothesis, leaving us with a known alternative. Absent meaningfully large samples selected at random, we are left with either the need to control for a

great many additional variables or the existence of a great many alternative hypotheses that could account for the observed result. In applied research, seldom do researchers have sufficient control over samples to meet this criterion. Employees may refuse to participate or may abandon their participation for any of a long list of reasons, ranging from hostility toward the employing organization to fear of revealing some negative information about themselves. Research participants may not be available for data collection when we need them for a variety of reasons, such as operational demands, illness, or simply vacation time. More often than not, we are constrained to work with the sample that is available.

Finally, I would note, though somewhat in passing, what has long been termed the *criterion problem*. Binning and Barrett (1989) make the point that “*a job performance domain is a construct* [emphasis in original]” (p. 480), though we often give too little attention to the development of measures of it for the purposes of validation research (Schmitt & Landy, 1993). Guion (1987) has commented that he wished he could discern a movement away from the use of supervisory ratings as criteria, but he could not. I suspect that most industrial and organizational psychologists would acknowledge the many problems associated with this most common of criteria. Their continued use, however, is almost certain for three reasons: (1) there are many jobs for which it is extremely difficult, if not impossible, to identify relevant objective performance measures at the individual employee level, (2) the costs and practical problems associated with obtaining such criteria can be virtually insurmountable, and (3) objective measures sometimes suffer from psychometric issues that render them inappropriate for use as criteria. So, it seems likely that we will continue to use the flawed, but available, evaluations of individuals by supervisors.

Business

Although I have treated business issues as constraints on our research, it is important that we not forget that among the reasons we seek to improve selection into organizations is to achieve improvements in productivity. Organization leaders rightly ask, “How will doing this research improve our competitive position or further our mission?” It is thus incumbent upon us to respond to or-

organizations' needs to show relevance and utility for the research we propose. Our failure to do so may constrain both the nature and the extent of what we will be allowed to do.

It is essential never to underestimate the implications of costs and budgets on the research we propose and conduct for organizations. Smaller organizations and those in businesses with lower profit margins must be concerned about every dollar expended. In larger, very profitable organizations, it is easy to fall into the trap of comparing the cost of selection research to the organization's revenues or what it spends on product research, but these are seldom the comparisons that matter. In actuality, the costs of research tend to be viewed relative to the discretionary budgets allocated to human resource functions, which are usually much smaller. Public sector organizations are far from immune to these pressures; indeed, by law most are required to engage in competitive bidding that emphasizes costs along with technical competence. Even when researchers conduct analyses to demonstrate the utility of selection programs, these analyses are often presented after the research is complete to justify the implementation of a testing program. In other words, we seldom conduct and present utility arguments (even when we have the data) as part of our proposals to justify conducting the research in advance, and even when we do, these are frequently dismissed as not credible.

And the costs for the research we do can be high. One example is the cost for data collection alone. It is cost prohibitive for many organizations to collect data from a number of far-flung but sparsely staffed facilities. Moreover, for all organizations, data collection usually means lost productivity due to employees' absence from their jobs or additional costs incurred for overtime. This issue accounts in some substantial part for both the small sample sizes that are obtained and the tendency of many researchers to resort to samples of convenience.

As scientists, we are in no position to make promises or guarantees regarding the outcome of our research, and even the most carefully conducted validation study may not produce the desired results. Moreover, in order to gain acceptance for this complex and expensive undertaking, all too often the arguments for conducting validation research lean heavily on legal liability, rather than on the intrinsic value of the knowledge to be gained.

As the world economy has become ever more competitive, many businesses (and public sector entities) have adopted the spirit of NASA's one-time mantra of "faster, better, cheaper." It is up to us, then, to offer both (1) better explanations of the need for and value of the research we need to conduct to investigate the validity of our selection procedures and (2) creative and substantive alternative strategies for developing and leveraging this evidence.

The fact that organizations will not sit still or even slow down while I/O psychologists conduct selection research means that that research may be constrained by ongoing events. Changes in the business environment at the macro-economic level may change both budgets and the perceived need for selection of new employees at all. The research that was viewed as valuable and necessary previously may become the cost-saving budget cut in a surprisingly short time frame. Mergers, acquisitions, sales, and purchases of businesses may make the research moot or cast doubt on its continued relevance or applicability.

The research that is acceptable or even allowed may be limited or changed by the intervention of third-party influences, including bargaining units and the general public (typically as represented by the media). It is not usually the case that union contracts place pre-employment selection under their purview (though some may). However, research that involves the participation of bargaining unit members (to provide job analysis information, serve as SMEs, or respond to surveys or tests) will require the cooperation (or at least nonresistance) of union leadership. If the measures to be researched are for promotion or certification, one can expect substantial involvement by these leaders and, in some cases, considerable resistance. Such resistance may change the level of cooperation of participants, change their psychological set while participating, alter the composition of the available sample, and possibly render the research infeasible. Engaging in validation research under the pressure of public scrutiny (and political oversight in the public sector) can have equally daunting effects. Certainly, no flaw in the research—even if unavoidable—will go unnoticed, and failure to produce "good" results (that is, significant correlations, low subgroup differences, and so on) will be attributed to failures of the research or even of the researcher. Finally, the external influence of ongoing or pending litigation (or

enforcement agency actions) may make organization leaders reluctant to pursue research. In some cases, the existence of previous litigation may impose specific or general constraints on the nature, type, and expectations for validation research.

A final “business” issue lies in a bias that human resource work is basically pretty easy stuff. This bias has two implications. First, it leads to a tendency to assume the veracity of personal theories of selection. Successful managers and business leaders have histories of making reasonably good decisions, and they often develop idiosyncratic models for how best to select the people they need for their organizations. These personal theories account in part for the endurance and ubiquity of traditional unstructured interviews despite years of research demonstrating their flaws. It is because of such theories, too, that managers so often cannot (or choose not to) understand our “unreasonable” need to conduct validation research when it is obvious to them that the test must be measuring things important to the job (such as mechanical or mathematical ability). Second, the bias about the simplicity of human resource functions leads managers to a belief that they, too, are experts in this area of the business. From their perspective, we are making it harder than it needs to be and dressing it up in unnecessary complexity and jargon. This bias may result in business leaders being susceptible to each new fad in testing that shows up on the Internet. It also accounts for their questioning our research demands: you don’t need a sample that big; surely the fifty people in our main plant will be more than sufficient.

Legal

In discussing business constraints, I mentioned litigation in passing. These issues have, however, become ever more prominent for I/O practitioners. As I noted previously, the existence of ongoing litigation or investigations may make business leaders reluctant to undertake or continue validation research. In addition, organizations may enter into agreements with enforcement agencies or have conditions imposed on them by courts that constrain their selection practices and limit the alternatives that they may consider.

More commonly, however, it is not the actuality of legal or enforcement action that impacts validation research in organizations.

Rather, it is concerns about the risk of such events, often advanced or supported by internal or external counsel, that impact our research opportunities and methods. The actual risks of litigation are a function of (1) the extent of exposure (in terms of amount of use of a selection procedure and the level of disparate impact resulting from that use) and (2) the extent to which there is evidence supporting its use (defensibility) (McPhail, 2005). Organizations and their legal counsel may impose requirements for either that cannot be achieved given the constraints imposed by the situation. In these situations, alternative and supplementary evidentiary sources may be necessary to meet the criteria.

Temporal

The business world moves quickly, and, more and more, the public sector world does so as well. Elapsed time for completion of validation research can become a serious concern for researchers. From the business perspective, there is a need for results to be available quickly. If business leaders are convinced that there is competitive advantage in improving selection, delay in realizing that advantage may be perceived as a business loss. Further, validation research is often not undertaken until there is an urgent “felt need,” thus increasing the pressure to take “short cuts” to complete the research in as short a time as possible. Research that takes an extended period of time may fall from a priority to a nonessential; it may lose its sponsor or its budget; and it may lose its relevance to a crucial event—such as a plant opening, a new acquisition, or a hiring window—from which the felt need originally grew.

Another aspect of time constraint is availability of time from internal resources, especially employees to serve as contributors or participants in the validation research. As noted earlier, this can be a cost issue, but beyond the costs, leaner organizations have fewer employees available and the impact of their absence from their jobs is greater. This constraint applies to an even greater extent to those supervisors who have increasing numbers of employees within their spans of control.

Organizational cycles also may impose significant time constraints in terms of both operational dynamics and availability of employees. For example, plant maintenance shutdowns, seasonal

work variability, and financial cycles may all impact an organization's willingness and ability to support and participate in validation research.

Organizational and Logistical

A long list of organizational and logistical issues arises to constrain research efforts. Of particular importance is the primacy of operational need. We cannot demand access to research participants in conflict with business goals, safety requirements, or time urgent organizational activities (such as provision of emergency services or completion of required maintenance activities).

As noted previously, validation research requires the active cooperation and participation of employees. Their fear of organizational consequences from the use of test and performance data in ways inimical to their best interests (whether justified or not) can have at least three effects: (1) nonrandom, nonrepresentative samples, (2) smaller samples, and (3) undetectably bad data (or increased error variance) from inattention or intentional undermining of the research. Finally, organizational structure, operating procedures, and schedules may impose situational constraints that preclude standard scientific research designs.

Alternative Strategies

Despite the organizational advantages (aside from the scientific value) that accrue from accurate information about the usefulness and job-relatedness of selection procedures, in many cases, employers engage in validation research solely to comply with legal obligations. The result has too often been hurriedly conducted, small-sample research barely sufficient to meet those minimum requirements imposed by the *Uniform Guidelines*. Because of resource constraints, many employers turn to tests misleadingly advertised as being "pre-validated," "EEOC approved," or "self-validating," or to selection procedures of questionable utility.

I/O psychology has offered a number of alternative strategies that may provide time and cost advantages and in some cases may improve on the accuracy of results obtained from small, limited samples. These alternatives are not widely known, and in some

cases they are misapplied or applied in inappropriate situations. The objective of this volume is to present these alternative strategies in sufficient detail and with appropriate cautions to allow practitioners greater access to their meaningful application.

Conclusion

When faced with situations like those described at the beginning of this chapter, too often I/O psychology has responded to organizations' requests for selection by saying, "No, you can't do that," rather than "Here is what we can do in this situation." Frequently, when faced with the inevitable contingencies of organizational reality (trade-offs in budgets, available personnel, and competing human resource objectives), decision makers have come to believe that I/O psychologists are insensitive to their needs. This volume provides practitioners with viable alternatives that allow them to address in a professionally acceptable manner a broader array of problems with a more sophisticated set of tools.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association.
- American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: APA.
- Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Mahwah, NJ: Erlbaum.
- Arvey, R. D. (1992). Constructs and construct validation: Definitions and issues. *Human Performance*, 5(1 & 2), 59–69.
- Arvey, R. D., & Faley, R. H. (1988). *Fairness in selecting employees* (2nd ed.). Reading, MA: Addison-Wesley.
- Ayer, A. J. (1935/1946/1952). *Language, truth and logic*. New York: Dover.
- Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, 74(3), 478–494.
- Blackburn, S. (1996). *Oxford dictionary of philosophy*. New York: Oxford University Press.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.

- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally & Company. (Reprinted from *Handbook of research on teaching*, by the American Educational Research Association, 1963.)
- Cohen, J. (1994/1997). The earth is round ($p < .05$). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 21–35). Mahwah, NJ: Erlbaum. (Reprinted from The earth is round ($p < .05$), J. Cohen, 1994, *American Psychologist*, 49(12), pp. 997–1003.)
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.
- Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (proceedings of a symposium in honor of Lloyd G. Humphreys). Urbana, IL: University of Illinois Press.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- Equal Employment Opportunity Commission. (1966). *Guidelines on employee testing procedures*. Washington, DC: Equal Employment Opportunity Commission.
- Equal Employment Opportunity Commission. (1970). *Guidelines on employee selection procedures*. *Federal Register*, 35(149), 12333–12336.
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). *Uniform guidelines on employee selection procedures*. *Federal Register*, 43(166), 38295–38309.
- Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist*, 27(2), 197–222.
- Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, 11(3), 385–398.
- Guion, R. M. (1987). Changing views for personnel selection research. *Personnel Psychology*, 40, 199–213.
- Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology*, 123(3), 207–215.

- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.
- Landon, T. E., & Arvey, R. D. (2006). Practical construct validation for personnel selection. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. xx–xx). San Francisco: Jossey-Bass.
- Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, *41*(11), 1183–1192.
- Lees-Haley, P. R. (1996). Alice in validityland, or the dangerous consequences of consequential validity. *American Psychologist*, *51*(9), 981–983.
- Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, 14–16.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.
- Markus, K. A. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible? *Social Indicators Research*, *45*, 7–34.
- McCormick, E. J. (1959). The development of processes for indirect or synthetic validity: III. Application of job analysis to indirect validity. A symposium. *Personnel Psychology*, *12*, 402–413.
- McPhail, S. M. (2005). Auditing selection processes: Application of a risk assessment model. *The Psychologist-Manager Journal*, *8*(2), 205–221.
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, Summer, 16–18.
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Mahwah, NJ: Erlbaum.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Popham, W. J. (1997). Consequential Validity: Right concern—Wrong concept. *Educational Measurement: Issues and Practice*, Summer, 9–13.
- Popper, K. (1935/1959/2002). *Logic of scientific discovery* (15th ed.). New York: Taylor & Francis.
- Rindskopf, D. M. (1997). Testing “small,” not null, hypotheses: Classical and Bayesian approaches. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 319–332). Mahwah, NJ: Erlbaum.

- Sackett, P. R., & Arvey, R. D. (1993). Selection in small N settings. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 418–447). San Francisco: Jossey-Bass.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274.
- Schmitt, N. (1989). Construct validity in personnel selection. In B. J. Fal-lon, H. P. Pfister, & J. Brebner (Eds.), *Advances in industrial organi-zational psychology* (pp. 331–341). New York: Elsevier Science.
- Schmitt, N., & Landy, F. J. (1993). The concept of validity. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 275–309). San Francisco: Jossey-Bass.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Edu-cation*, *19*, 405–450.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, Summer, 5–8, 13, 24.
- Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures*. Bowling Green, OH: Society for Industrial and Organizational Psychology.
- Thornton, S. (2005). Karl Popper. In E. N. Zalta (Ed.), *The Stanford ency-clopedia of philosophy* (Summer 2005 ed.). Retrieved from <http://plato.stanford.edu/archives/sum2005/entries/popper/>
- Wainer, H., & Braun, H. I. (Eds.). (1988). *Test validity*. Mahwah, NJ: Erlbaum.
- Zimiles, H. (1996). Rethinking the validity of psychological assessment. *American Psychologist*, *51*(9), 980–981.

