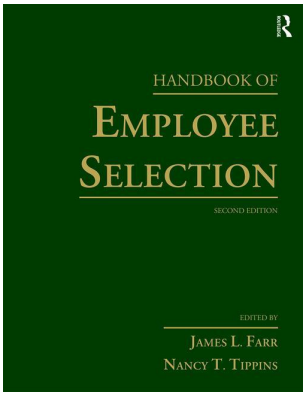

Handbook of Employee Selection

James L. Farr, Nancy T. Tippins, Walter C. Borman, David Chan, Michael D. Coovert, Rick Jacobs, P. Richard Jeanneret, Jerard F. Kehoe, Filip Lievens, S. Morton McPhail, Kevin R. Murphy, Robert E. Ployhart, Elaine D. Pulakos, Douglas H. Reynolds, Ann Marie Ryan, Neal Schmitt, Benjamin Schneider

Validation Strategies for Primary Studies

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-2
Neal W. Schmitt, John D. Arnold, Levi Nieminen
Published online on: 22 Mar 2017

2

Validation Strategies for Primary Studies

NEAL W. SCHMITT, JOHN D. ARNOLD, AND LEVI NIEMINEN

NATURE OF VALIDITY

Most early applications of the use of tests as decision-making tools in the selection of personnel in work organizations involved a validation model in which the scores on tests were correlated with some measure or rating of job performance, such as the studies of salespersons by Scott (1915) and streetcar motormen by Thorndike (1911). This view of validity was reinforced in books by Hull (1928) and Viteles (1932). Subsequent reviews by Ghiselli (1966, 1973) were similarly focused on what was by then known as criterion-related validity.

During this time, there was a recognition that tests could and should be based on other logical arguments as well. *Standards for Educational and Psychological Tests* (American Psychological Association [APA], American Educational Research Association [AERA], and National Council on Measurement in Education [NCME], 1954) identified four aspects of validity evidence: content, predictive, concurrent, and construct. With time, the predictive and concurrent aspects of validity became seen as simply different research designs, the purpose of which was to establish a predictor-criterion relationship; hence, they became known as criterion-related validity. Content and construct validation were seen as alternate methods by which one could validate and defend the use of test scores in decision making. A much broader view of the nature of validity is accepted today, and in general it is seen as the degree to which the inferences we draw from a set of test scores about job performance are accurate.

Subsequent separation of approaches to validity (content, construct, and criterion-related) produced numerous problems, not the least of which was the notion that there were times when one approach was to be preferred over another or that there were different acceptable standards by which these different aspects of validity were to be judged. Most important, however, was the realization on the part of measurement scholars that all were aspects of construct validity—the theoretical reasonableness of our explanations of job behavior. There was a realization that the inferences we derive from test scores was central to all validation work. Content validity and the practices usually associated with it were recognized as desirable practices in the development of any test. Careful consideration of the "theory" and hypotheses that underlie our conceptualization of performance and how the constructs central to job performance are represented in our tests is always important and unifying insofar as our validation efforts are concerned. Traditional criterion-related research represents one type of evidence that can be collected to confirm/disconfirm these hypotheses. This "unitarian" approach to validity was strongly argued in the 1985 *Standards* and has been incorporated in the 1999 and 2014 versions of the *Standards* (AERA, APA, & NCME, 1999; 2014). In all instances, evidence from multiple studies or sources is desirable.

Different Approaches to the Collection of Data About Validity

Validity, as defined in the most recent version of the *Standards* (2014), is "the degree to which evidence and theory support the interpretation of test scores for proposed uses of the test" (p. 11). The user must state explicitly what interpretations are to be derived from a set of test scores, including the nature of the construct thought to be measured. The document goes on to describe a variety of evidence that can support such an interpretation.

Content Evidence

An evaluation of test themes, wording, item format, tasks, and administrative guidelines all constitute the "content" of a test, and a careful logical or empirical analysis of the relationship of this content to the construct measured as well as expert judgments about the representativeness of the items to the construct measured supports validity. The evidence that a measure is content valid usually takes the form of an analysis by subject matter experts that describes a linkage between the test content and the content of a job. Perhaps most stringent is the view that a test is content valid if it is a "representative sample of the tasks, behaviors, or knowledge drawn from that domain" (*Principles*, 1987, p. 19), meaning the job domain. A more liberal approach to content validity is expressed in the 2003 version of the *Principles*. That is, a test is content valid if there is evidence that the test was designed explicitly as a sample of the "important work behaviors, activities and/or worker KSAOs necessary for performance on the job or in job training" (p. 21). Also, content validity evidence may include "logical or empirical analyses that compare the adequacy of the match between test content and work content, worker requirements, or outcomes of the job" (p. 6, *Principles*, 2003). The role of content evidence in the validation process continues to be controversial among professionals in the field, as evidenced by the paper authored by Murphy (2009) and the responses to his paper in the same issue of *Industrial and Organizational Psychology*. Murphy stated what some in this series of papers thought was old news; namely, that evidence that job content and test content were highly similar was not related to criterion-related validity. Responses reflected a variety of views as to the nature of content validity and the notion that there was no reason the sets of evidence should be related.

In a typical content validity effort, subject matter experts provide judgments that link knowledge, skills, abilities, and other characteristics (KSAOs) to specific job elements (i.e., a KSAO is required to perform a part of the job adequately) and link KSAOs to test items or subtests (i.e., responses to a test item provide information about the level of a test taker's KSAO). An effort is then made to assess the communality between these two lists of KSAOs and conclude that the communality is (not) sufficient to support the inference that people who do well on the test will also do well on the job.

A concern in some instances is the degree to which the results of a content validity study conducted in one context can be used to support inferences about job performance in another situation. Perhaps the most common rationale for such generalization is the notion that the new or local setting is similar to that in which the content validity study was done; that is, characteristics of the applicant, the predictor and criterion constructs, and other important aspects of the two situations (that of the original content validation study and that of the situation to which results are to be generalized). The arguments that the work components of the two situations are the same must be clear and persuasive.

Response Processes

Validity evidence can also take the form of an examination of the response processes involved in responding to an item. For example, in evaluating the capabilities of an applicant for a mechanical job, we might ask the person to read a set of instructions on how to operate a

piece of equipment and then ask the applicant to demonstrate the use of the equipment. Because the equipment is used on the job, it would seem valid, but suppose we also find that test scores are highly related to examinees' vocabulary level. We would then want to know if vocabulary is necessary to learn how to use this equipment on the job and, depending on the answer to that question, we may want to revise the test. It is rare, in our experience at least, that the similarity of response processes across tests and criteria are presented as the sole validity support, but they are often inherent in what is more likely to be termed content validity or transportability arguments (i.e., transporting a validity claim from one situation to another).

Internal Structure of the Test

Yet a third piece of evidence might be to collect data regarding the internal structure of a test. We would examine the degree to which different items in a test (or responses to an interview) yield correlated results and whether items designed to measure one construct can be differentiated from items written to assess a different construct. Researchers interested in these questions use item means, standard deviations, and intercorrelations as well as exploratory and confirmatory analyses to evaluate hypotheses about the nature of the constructs measured by a test. When these data confirm the hypothesized nature of the constructs measured by the test and those constructs are deemed to underlie worker performance, there is support for the predictive inference (i.e., test scores predicts job performance).

Criterion-Related Evidence

Similar to looking at the internal structure of a test, researchers can also examine its external validity by correlating the test results with job performance measures. Validity in the personnel selection area has been almost synonymous with the examination of the relationship between test scores and job performance measures, most often referred to as criterion-related validity. Because there is a large body of primary studies of many job performance-test relationships, one can also examine the extent to which tests of similar constructs are related to job performance and generalize in a way that supports the validity of a new measure or an existing measure in a new context. These are studies of validity generalization, which we will discuss in more depth in the Validity Generalization section. It should be noted that without primary studies of criterion-related validity, there can be no validity generalization studies, and without recent studies of criterion-related validity, we cannot assess newer developments in testing technology or criterion development using meta-analyses. Likewise, validity transportability and synthetic validity (see discussion of synthetic validity below) support for the predictive hypothesis underlying the use of tests are impossible without primary studies of criterion-related validity.

In practice, criterion-related validity studies are often criticized for failing to adequately address validity issues surrounding the criterion measure(s) used. The relative lack of scientific scrutiny focused on criteria, termed the "criterion problem" (Austin & Villanova, 1992), has been a topic of discussion among personnel psychologists for years (Dunnette, 1963; Fiske, 1951; Guion, 1961). Universal to these discussions is the call for more rigorous validation evidence with respect to the criteria that are used. Binning and Barrett (1989) outlined this task, underscoring two interrelated goals for the validation researcher. First, they suggested that the selection of criteria should be rooted in job analysis to the same extent that selection of predictors traditionally are (i.e., more attention to rigorous "criterion development"). Other considerations relevant to the tasks of criterion development and validation include the use of "hard" or objective criteria versus more proximal behaviors that lead to these outcomes (Thayer, 1992), use of multiple relevant criteria as opposed to a single overall criterion (Campbell, McCloy, Oppler, & Sager, 1993; Dunnette, 1963), and the possibility

that criteria are dynamic (i.e., change over time for employees as a function of how long they have been on the job) (Barrett, Caldwell, & Alexander, 1985). Second, researchers should be concerned with demonstrating evidence of construct-related validity for the criterion. Investigators must specify the latent dimensions that underlay the content of their criterion measures. This involves expansion of the nomological network to include inferences that link the criterion measure(s) to constructs in the performance domain (e.g., by demonstrating that criterion measures are neither contaminated nor deficient with respect to their coverage of the intended constructs in the performance domain) and link constructs in the performance domain to job demands that require specific ability or motivational constructs (e.g., by demonstrating through job analysis that constructs in the performance domain are organizationally meaningful). Campbell and his colleagues (e.g., Campbell, McCloy, Oppler, & Sager, 1993) have repeatedly emphasized the importance of the nature of criteria or performance constructs. These authors make the somewhat obvious, although often overlooked, point that performance should be defined as behavior ("what people actually do and can be observed"); the products of one's behavior, or what are often called "hard criteria," are only indirectly the result of one's behavior and may be influenced by other factors that are not attributable to an individual job incumbent. Further, we may consider relatively short-term or proximal criteria or distal criteria, such as the impact of one's career on some field of interest. Any specification of a performance or criterion domain must also consider the impact of time (Ackerman, 1989; Henry & Hulin, 1989). In any study of performance, these various factors must be carefully considered when one decides on the nature of the performance constructs and actual operationalizations of the underlying constructs and how those measures might or might not be related to measures of other constructs in the domain of interest.

Use of a criterion-related strategy makes a special set of methodological and statistical approaches relevant. Power analysis is a useful framework for interrelating the concepts of statistical significance, effect size, sample size, and reliability (Cohen, 1988) and has design and evaluation implications for the statistical relationships sought in criterion-related studies. For instance, the sample size needed to demonstrate a statistically significant predictor-criterion relationship decreases as the magnitude of the relationship that exists between predictor and criterion (i.e., effect size) increases. Sussman and Robertson (1986), in their assessment of various predictive and concurrent validation designs, found that those strategies that allowed larger sample sizes gained a trivial increment in power. This suggests that, as long as sample sizes can support the use of a criterion-related design, further attention toward increasing N may not reap large benefits. Other factors affecting power include the interrelatedness and number of predictors used, such that the addition of nonredundant predictors increases power (Cascio, Valenzi, & Silbey, 1978). The reliability of the predictors and criteria and the decision criteria used for inferring that a relationship is nonzero (i.e., the confidence interval around the estimate of effect size is not zero) also impact power.

By incorporating power analysis in validation design, researchers can increase the likelihood that relationships relevant to key inferences will be tested with sufficient sample size upon which to have confidence in the results. However, from a scientific standpoint, the importance of demonstrating that predictor-criterion relationships are statistically significant may be overstated, given that relationships, which may not be practically meaningful, can reach statistical significance with large enough sample sizes. For instance, a statistically significant relationship, in which a test accounts for less than 5% or a relatively small portion of the variance in job performance, is not unequivocal support for the test's use. This is especially evident when there is reason to suggest that other available tests could do a better job predicting performance. Further, rather than rely on statistical significance tests, an argument about the practical utility of the information about test takers' KSAOs is most relevant and important (see discussion of utility later in this chapter).

Operationally, there are several other important considerations in criterion-related research (e.g., job analyses that support the relevance of predictor and criterion constructs and the quality of the measures of each set of constructs). However, those concerns are addressed repeatedly in textbooks (e.g., Guion, 1998; Ployhart, Schneider, & Schmitt, 2006). In the next section of this chapter, we address a very important concern that is rarely discussed.

Transportability of Validity

Another factor that can affect the extent of the local validation effort that is required is the availability of existing validation research. The *Principles* describes three related validation strategies that can be used as alternatives to conducting traditional local validation studies or to support the conclusions drawn at the primary study level. First, "transportability" of validity evidence involves applying validity evidence from one selection scenario to another, on the basis that the two contexts are judged to be sufficiently similar. Specifically, the *Principles* note that researchers should be concerned with assessing similarity in terms of characteristics [e.g., the knowledge, skills, and abilities (or KSAs) needed to perform the job in each context], job tasks and content, applicant pool characteristics, or other factors that would limit generalizability across the two contexts (e.g., cultural differences). Assessing similarity in this manner usually requires that researchers conduct a job analysis or rely on existing job analysis materials combined with their own professional expertise and sound judgment—and documenting carefully all procedures used to inform the decision.

Synthetic Validity

Synthetic validity is a process in which validity for a test battery is "synthesized" from evidence of multiple predictor–job component relationships (Peterson, Wise, Arabian, & Hoffman, 2001; Scherbaum, 2005). Job analysis is used to understand the various components that make up a particular job, and then predictor–job component relationships are collected for all available jobs with shared components. Because evidence can be drawn from other jobs besides the focal job, synthetic validity may be a particularly useful strategy for organizations that have too few incumbents performing the focal job to reach adequate sample sizes for a traditional criterion-related study (Scherbaum, 2005). Transportability and synthetic validity are similar notions; in transportability, one is taking the entire results of a validation study to justify use of a test or test battery in a new situation. In synthetic validity, one is taking the results of multiple different studies on different constructs to justify their use in a new situation in which the various constructs are deemed important to successful job performance.

An excellent example and evaluation of a synthetic validation effort is provided by Johnson and Carter (2010). Job analysis data in a large organization provided evidence for 11 job families and 27 job components. Twelve tests were developed to predict performance on these job components. Test scores and performance data on the job components were collected from 1,926 incumbents. A test composite for each job component was created, and a test battery was chosen for each job family based on relevant job components. Synthetic validity coefficients computed on each battery compared favorably with traditional validity coefficients computed within those job families for which adequate sample sizes were available.

Validity Generalization

Validity generalization involves using meta-analytic findings to support the conclusion that predictor-criterion validity evidence can be generalized across situations. Like transportability strategies, meta-analytic findings provide researchers with outside evidence to support inferences in a local context. The argument for validity generalization on the basis of meta-analyses is that some selection tests, such as cognitive ability tests (Ones, Viswesvaran, & Dilchert, 2005), are valid across selection contexts. Thus, the implication is that with validity generalization strategies, unlike transportability, in-depth job analyses or qualitative studies of the local organizational context are unnecessary. In support of this assertion, Schmidt and Hunter and colleagues (for review, see Schmidt & Hunter, 2003) have argued that between-study variability in validity coefficients can be largely attributed to statistical artifacts, such as range restriction, unreliability, or sampling error. However, caution is warranted to the extent that meta-analyses have identified

substantive moderators, or in the presence of strong theory indicating that some variable may moderate the magnitude of validity. Further, with regard to generalization across contexts, inferences drawn from meta-analytic findings are limited to the contexts of those studies included in the meta-analysis (LeBreton et al., Chapter 4 of this volume).

When local criterion-related studies of some relationship are conducted, meta-analytic estimates of the same relationship may be used as Bayesian priors to estimate the degree to which a meta-analytic estimate should be modified by the new local evidence. Schmidt and Hunter (1977) discussed this possibility in their original validity generalization study, but few researchers have recognized or used meta-analytic evidence in this fashion. However, recognition of the utility of such an approach to science appears to be increasing (e.g., Zyphur, Oswald, & Rupp, 2015). At minimum, meta-analytic findings should be referenced in test development and can be used to supplement evidence at the local level, either via theoretical or statistical means (Newman, Jacobs, & Bartram, 2007). The argument for more direct use of validity generalization strategies is dependent on the strength of the meta-analytic findings and in some cases may mean that local validation efforts are unnecessary or even misleading (e.g., due to small sample sizes). Nevertheless, the legal defensibility of the selection procedure may necessitate a local validation study.

Evidence Regarding the Consequences of Test Use

Finally, and somewhat controversially among industrial-organizational (I-O) psychologists, the *Standards* (1999, 2014) also suggest that researchers examine the intended and unintended consequences of test use to make decisions. This evidence is referred to as *consequential validity* (Messick, 1998). The consequences of most concern are the degree to which use of test scores results in disproportionate hiring of one or more subgroups (e.g., gender, race, disabled).

Finally, some I-O psychologists have also noted that the traditional separation of reliability and validity concepts may be inadequate (Lance, Foster, Gentry, & Thoresen, 2004; Lance, Foster, Nemeth, Gentry, & Drollinger, 2007; Murphy & DeShon, 2000). Technology also affords the opportunity to make the traditional one-time criterion-related validity study an ongoing effort in which the accumulation of predictor and criterion data can be collected and aggregated across time and organizations.

VALIDATION IN DIFFERENT CONTEXTS

This chapter discusses validation largely within the context of personnel selection. This is the most common application of the various approaches to validation. It is also the most straightforward example of how validation approaches can be applied.

There is a wide range of contexts in which the validation of measures is desirable; however, organizations should, for example, ensure they are using "validated" tools and processes in their performance management systems, in their assessments of training and development outcomes, in their promotion and succession planning processes. In some instances, there should also be validity evidence for the use of survey measures (e.g., when the use of the measure is predicated on the notion that it measures some well-known construct and a scientific body of research is used to defend the use of the survey measure).

Each of these circumstances is associated with its own set of challenges as the researcher designs an appropriate validation study. However, the design of the well-constructed study by necessity will follow the same logic as will be discussed for the personnel selection context. Following this logic, the studies should be structured to include the following three elements:

1. *Job analysis*. The foundation of validation in employment settings always involves the development of a clear understanding of job and organizational requirements. For example, for promotion purposes these would be the requirements of the target job(s) into which a person might be promoted. For training and development purposes, these would be the meaningful outcomes in terms of on-the-job performance that are the focus of the training/development efforts.

2. *Systematic development*. As measures are developed, they need to follow an architecture that is firmly grounded in the results of the job analysis. As the development of the measures is planned and as the tools are being constructed, activities need to be focused on ensuring that the measures are carefully targeted to address the intended constructs.
3. *Independent verification*. Once the measures are developed, they need to be subjected to independent verification that they measure the intended constructs. At times, this can involve statistical studies to determine whether the measures exhibit expected relationships with other independent measures (e.g., Does the 360-degree assessment of leadership behavior correlate with an independent interview panel's judgment of a leader's behavior?). The independent verification is often derived from structured expert reviews of the measures that are conducted prior to implementation. Regardless of the method, this "independent verification" is a necessary aspect of verifying the validity of a measure.

Strong Versus Weak Inferences About Validity

Given that validation is a process of collecting evidence to support inferences derived from test scores (e.g., that a person will perform effectively on a job), the confidence with which inferences are made is a function of the strength of the evidence collected. Gathering stronger evidence of validity almost always necessitates increased effort, resources, and/or costs (e.g., to gain larger sample sizes or expand the breadth of the criterion measures). Thus, a key decision for researchers designing primary validation studies involves determining how to optimize the strength of the study (assurance that inferences are correct) within the bounds of certain practical limitations and organizational realities. Situations may vary in terms of the extent to which feasibility drives the researcher's choice among validation strategies. In some cases, situational limitations may be the primary determinant of the validation strategies available to researchers. For example, for situations in which adequately powered sample sizes cannot be achieved, validation efforts may require use of synthetic validity strategies (Scherbaum, 2005), transporting validity evidence from another context that is judged to be sufficiently similar (Gibson & Caplinger, 2007), generalizing validity across jobs or job families on the basis of meta-analytic findings (McDaniel, 2007; Rothstein, 1992), or relying on evidence and judgments that the content of the selection procedures is sufficiently similar to job tasks and/or the KSAOs required to support their use in decision making. Other factors noted by the *Principles* that may limit the feasibility of certain validation strategies include unavailability of criterion data, inaccessibility to subject matter experts (SMEs), as might be the case when consulting SMEs would compromise test security, dynamic working conditions such that the target job is changing or does not yet exist, and time and/or money.

Given the need to balance several competing demands (e.g., issues of feasibility limiting the approach that can be taken versus upholding high standards of professionalism and providing strong evidence to support key inferences), it is essential that researchers understand the various factors that have potentially important implications for the strength of evidence that is required in a given validation scenario. In other words, part of the decision process, with regard to planning and implementing a validation strategy, is a consideration of how strong the evidence in support of key inferences ought to be. The basic assumption here is that different situations warrant different strategies along several dimensions (Sussman & Robertson, 1986), one of which has to do with the strength of evidence needed in support of inferences. Rather, all validation studies and selection practices should aspire to the ethical and professional guidelines offered in the *Principles*, which means using sound methods rooted in scientific evidence and exhibiting high standards of quality. However, the *Principles'* guidelines are not formulaic to the exclusion of professional judgment, nor are their applications invariant across circumstances. In the following paragraphs, several factors are identified that have potential implications for the strength of the evidence needed by a local validation effort.

Situational Factors Influencing the Strength of Evidence Needed

Although it is beyond the scope of this chapter to describe in full detail the legal issues surrounding validation research and selection practice (see Chapters 28, 29, and 30, this volume,

for further discussions of legal issues), it would be difficult if not impossible to describe applied validation strategy without underscoring the influence of litigation or the prospect of litigation in the U.S. It is becoming almost cliché to state that, in circumstances in which there is a relatively high probability of litigation regarding selection practices, validation evidence is likely to function as a central part of defending selection practices. Indeed, much validation research is stimulated by litigation, whether post facto or in anticipation of lawsuits. Within this context, researchers make judgments regarding the potential for litigation and adapt their validation strategies accordingly. Numerous contextual factors contribute to the probability that litigation will occur. A primary example has to do with the type of selection instrument being validated and the potential for *adverse impact*, or the disproportionate rejection of identifiable subgroups. Tests that have historically resulted in adverse impact, such as tests of cognitive ability (Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997) or physical ability (Arvey, Nutting, & Landon, 1992; Hogan & Quigley, 1986), tend to promote more litigation, and researchers validating these instruments in a local context should anticipate this possibility. Similarly, selection instruments with low face validity (i.e., the test's job relevance is not easily discerned by test takers) are more likely to engender negative applicant reactions (Shotland, Alliger, & Sales, 1998), and decisions based on such tests may lead to applicant perceptions of unfairness (Cropanzano & Wright, 2003). In their review of the antecedents and consequences of employment discrimination, Goldman, Gutek, Stein, and Lewis (2006) identified employee perceptions of organizational and procedural justice as important antecedents of discrimination lawsuits. In addition to considering characteristics of the selection instrument(s) being validated, lawsuits over selection practice are more frequent in some industry (e.g., government) and job types (Terpstra & Kethley, 2002).

Researchers should also consider the implications and relative seriousness of hiring decisions that result in false positives or false-negative errors. A false positive is made by selecting an unqualified individual whose performance on the job will be low, whereas a false-negative error is made by rejecting a qualified individual whose performance on the job would have been high. Making an error of either type can be potentially costly to the organization. However, the relative impact of such errors can differ by occupation type and organizational context. For example, the negative impact of a false positive in high-risk occupations (e.g., nuclear power plant operator or air-traffic controller) or high-visibility occupations (e.g., Director of the Federal Emergency Management Agency [FEMA]) can be catastrophic, threaten the organization's existence, and so on (Picano, Williams, & Roland, 2006). Alternatively, for occupations that are associated with less risk, such that failure on the job does not have catastrophic consequences for the organization or larger society, or when organizations use probationary programs or other trial periods, the cost of false-positive errors may be relatively low. Although validation efforts in both situations would be concerned with selection errors and demonstrating that use of tests can reduce the occurrence and negative consequences of such errors, clearly there are some situations in which this would be more of a central focus of the validation effort. It is our contention that validating selection systems for high-risk occupations are a special circumstance warranting particularly "watertight" validation strategies in which strong evidence should be sought to support the inferences made. In these circumstances, a test with low validity (e.g., less than $r = .10$) might be used to make hiring decisions when the outcome of such decisions is critically important to organizational effectiveness, and decision makers would want to use any evidence available to reduce risk even if they are not predicting a large amount of variance.

In some circumstances, the cost of false negatives is more salient. For example, strong evidence of a test's validity may be warranted when an organization needs to fill a position or several positions, but applicants' test scores are below some acceptable standard, indicating that they are not fit to hire (i.e., predicted on-the-job performance is low or very low). In this case, the organization's decision to reject an applicant on the basis of his/her test scores would leave a position or several positions within the organization vacant, a costly mistake in the event that false-negative errors are present. Demonstrable evidence to support the test's validity would be needed to justify such a decision, and in essence, convince the organization that it is better off with a vacant position than putting the wrong person in the job. In these instances, one might want evidence of a larger test-criterion relationship (perhaps greater than $r = .30$) to warrant use of this test and the possible rejection of competent applicants.

The possibility of false negatives becomes a special concern when members of some subgroup(s) are selected less frequently than members of another subgroup. When unequal ratios of various subgroups are selected, the organization must be prepared to show that false negatives are not primarily of one group as opposed to another. When this is impossible, the legal and social costs can be very high.

Personnel psychologists have long been aware of the fact that the utility of selection systems increases as a function of selectivity, such that selection instruments even modestly related to important outcomes can have large payoffs when there are many applicants from which only a few are to be selected (Brogden, 1951, 1959). On the other hand, as selection ratios become extremely liberal, such that nearly all applicants are accepted, even selection instruments highly related to performance have less positive implications for utility. From a purely utilitarian perspective, it would seem logical that demonstrating test validity is less of an impetus when selection ratios are liberal (because even the best tests will have little effect) and more of an impetus when selection ratios are low.

In licensing examinations, this utility perspective takes a different form because the major purpose of these examinations is to protect the public from "injuries" related to incompetent practice. In this case, the license–no license decision point using test scores is usually set at a point that is judged to indicate "minimal competence." Depending on the service provided (e.g., hairdresser vs. surgeon), the cost of inappropriately licensing a person could be very different. On the other hand, certification examinations are usually oriented toward the identification of some special expertise in an area (e.g., pediatric dentistry or forensic photography); hence, a decision as to a score that would warrant certification might result in the rejection of larger numbers or proportions of examinees. The cost-benefit balance in this situation (assuming all are minimally competent) might accrue mostly to the individual receiving the certification in the form of greater earning power.

Design Considerations When Strong Evidence Is Needed

On the basis of the preceding discussion, situational factors can affect the feasibility and appropriateness of the validation models applied to a given selection context. Moreover, researchers should be particularly attuned to contextual variables that warrant an increased concern for demonstrating the strength of evidence collected and high levels of confidence in the inferences to be made. The validity strategies used reflect consideration of these contextual factors and others. The discussion that follows is focused on identifying a handful of actionable validation strategies to be considered by researchers when particularly strong evidence is needed.

Importance of the Nomological Net

Binning and Barrett (1989) offered a thorough conceptualization of the nomological network implicit in validity models (see Chapters 1 and 3 in this volume). Their model identifies multiple inferential pathways interrelating psychological constructs and their respective operational measures. Inferential pathways in the model are empirically testable using observed variables (e.g., linkages between operationalized measures of constructs and linkages between constructs and their operationalized measures). Others may be theoretically or rationally justified (e.g., construct-to-construct linkages) or tested using latent variable models, although these applications are relatively rare in personnel selection research (see Campbell, McHenry, & Wise, 1990, for an attempt to model job performance). Consistent with the unitarian conceptualization of validity, all validity efforts in a selection context are ultimately concerned with demonstrating that test scores predict future job performance, and each of the various inferential pathways represents sources or types of evidence to support this common inference. Binning and Barrett (1989, p. 482) described how "truncated" validation strategies often concentrate exclusively on demonstrating evidence for a single inferential pathway and as a result provide only partial support for

conclusions regarding test validity. A more cogent argument for validity is built upon demonstration of strong evidence for several inferential pathways within the nomological network. For example, in addition to demonstrating a statistical relationship between observed measures from the predictor and performance domain, as is commonly the main objective in criterion-related validity studies, researchers should provide evidence of the psychological constructs underlying job performance (as well as the predictor measures) and demonstrate that the criterion measure adequately samples constructs from the performance domain.

Criterion Concerns

Various concerns regarding the criterion used in validation research are enumerated above in the description of criterion-related evidence of test validity. These concerns are particularly important when there is a need for strong evidence of test validity and tests are evaluated by relating test scores to measures of job performance.

Multiple Inferences in Validation Research

Gathering evidence to support multiple inferences within a theoretically specified nomological network resembles a pattern-matching approach. The advantage of pattern-matching research strategies is that stronger support for a theory can be gained when complex patterns of observed results match those that are theoretically expected (Davis, 1989). Logically, it would be less likely that a complex pattern of results would be observed simply because of chance. For example, when high scores on a test of empathy are expected to moderate the relationship between conscientiousness and the realization of performance goals and this pattern of expected relationships holds, it is unlikely due to chance. In addition, when experimental control of potentially confounding variables is not possible, pattern matching can be used to preempt alternative explanations for the observed relationships (i.e., threats to validity; Cook & Campbell, 1979).

A more extensive form of pattern matching involves the use of multiple studies, or research programs, to corroborate evidence of validity. Again, the logic is straightforward; stronger evidence is gained when a constellation of findings all lead to the same conclusion. Sussman and Robertson (1986) suggested that programs of research could be undertaken, "composed of multiple studies each utilizing a different design and aimed at collecting different types of evidence" (p. 467). Extending the rationale of the multi-trait multi-method (MTMM; Campbell & Fiske, 1959), convergent evidence across studies may indeed be stronger if gained through different research designs and methods. Landy's (1986) assertion that test validation is a form of hypothesis testing, and that judgments of validity are to be based on a "preponderance of evidence" (p. 1191; Guion, as cited in Landy, 1986), provides the context for consideration of research strategies such as quasi-experimental designs (Cook & Campbell, 1979) and program evaluation research (Strickland, 1979). Binning and Barrett (1989) presented a similar rationale by calling for "experimenting organizations" (p. 490) in which local validation research is treated as an ongoing and iterative process. Published research on use of multiple experiments or methods in a selection-validation context remains sparse to date.

CONCERNS ABOUT THE QUALITY OF THE DATA: CLEANING THE DATA

Once data have been collected, quality control techniques should be applied to ensure that the data are clean before proceeding to statistical analysis. Some basic quality control techniques include double-checking data for entry errors, spot checking for discrepancies between the electronic data and original data forms, inspecting data for out-of-range values and statistical outliers, and visually examining the data using graphical interfaces (e.g., scatterplots, histograms, stem-and-leaf plots). Special concern is warranted in scenarios in which multiple persons are

accessing and entering data or data sets from multiple researchers are to be merged. Although these recommendations may appear trite, they are often overlooked, and the consequence of erroneous data can be profound for the results of analyses and their interpretations.

A study by Maier (1988) illustrated, in stepwise fashion, the effects of data cleaning procedures on validity coefficients. Three stages of data cleaning were conducted, and the effects on correlations between the Armed Services Vocational Aptitude Battery (ASVAB) and subsequent performance on a work sample test for two military jobs (radio repairers and automotive mechanics) were observed. Selection was based on the experimental instrument (the ASVAB), and the work sample criterion tests were administered to incumbents in both occupations after some time had passed. In Phase 1 of the data cleaning process, the sample was made more homogenous for the radio repairers group by removing the data of some employees who received different or incomplete training before criterion data collection. In comparison to the total sample, the validity coefficient for the remaining, more representative group that had received complete training before criterion collection was decreased (from .28 to .09). The initial estimate had been inflated because of the partially trained group having scored low on the predictor and criterion.

In Phase 2, scores on the criterion measure (i.e., ratings from a single rater on a work sample) were standardized across raters. Significant differences among raters were attributed to different rating standards and not to group differences in ratees, such as experience, rank, or supervisor performance ratings. The raters were noncommissioned officers and did not receive extensive training in the rating task, so that differences among raters in judgmental standards were not unexpected. As a result, the validity coefficients for both jobs increased (radio repairers, from .09 to .18; automotive mechanics, from .17 to .24). In Phase 3, validity coefficients were corrected for range restriction, which again resulted in an increase in the observed validity coefficients (radio repairers, from .18 to .49; automotive mechanics, from .24 to .37). Maier noted that the final validity coefficients were within the expected range on the basis of previous studies.

The Maier (1988) study is illustrative of the large effect that data cleaning can have for attaining more accurate estimates of validity coefficients in a predictive design scenario. Several caveats are also evident, so that researchers can ensure that data cleaning procedures conducted on sound professional judgment are not perceived as data "fudging" and/or HARKing (Kerr, 1998). First, the cleaning procedures need to have a theoretical or rational basis. Researchers should document any decision criteria used and the substantive changes that are made. For example, researchers should record methods used for detecting and dealing with outliers. In addition, a strong case should be built in support of any decisions made. The researcher bears the burden of defending each alteration made to the data. For example, in the Maier study, the decision to standardize criterion data across raters (because raters were relatively untrained and used different rating standards) was supported by empirical evidence that ruled out several alternative explanations for the mean differences observed among raters. Perhaps the most serious problem is choosing which set of predictor-criterion relationships to report based on post hoc examination of the data. The best approach when using various corrections to observed data is to report both corrected and uncorrected values of data parameters.

Finally, missing data on some variables in many applied studies is common, and a whole science (Little & Rubin, 2002) has evolved around the imputation of missing values (i.e., estimating the value of a missing variable based on the level of other available data). Bayesian analyses applied to the problem of missing data have allowed for estimation of parameters and the confidence with which attributions about individual difference–performance relationships can be made. These techniques are not commonly used in personnel selection research, and we think they could be usefully applied in many instances.

MODES OF DECISION-MAKING AND THE IMPACT ON UTILITY AND ADVERSE IMPACT

If we have good-quality data, it still matters how we use those data in making decisions as to whether or not use of the test produces aggregated performance improvements. In this section, we will discuss the impact of various modes of decision making on two outcomes that

are of concern in most organizations: overall performance improvement or utility and adverse impact on some protected group defined as unequal proportions of selection across subgroups. Advancing both outcomes is often in conflict, especially when one uses cognitive ability tests to evaluate the ability of members of different racial groups or physical ability when evaluating male and female applicants for a position. Measures of some other constructs (e.g., mechanical ability) produce gender or race effects, but the subgroup differences that are largest and affect the most people are those associated with cognitive and physical ability constructs.

Top-Down Selection Using Test Scores

If a test has a demonstrable relationship to performance on a job, the optimal utility in terms of expected employee performance will occur when the organization selects the top-scoring persons on the test to fill its positions (Brown & Ghiselli, 1953). Expected performance is a direct linear function of the test score–performance relationship in the situation in which the top-scoring individuals are selected. However, use of tests in this fashion when it is possible will mean that lower-scoring subgroups will be less likely to be selected (Murphy, 1986). This conflict between maximization of expected organizational productivity and adverse impact is well known and has been quantified for different levels of subgroup mean differences in ability and selection ratios (Sackett & Wilk, 1994; Sackett, Schmitt, Ellingson, & Kabin, 2001; Schmidt, Mack, & Hunter, 1984). For social, legal, and political reasons, as well as long-term organizational viability in some contexts, the adverse impact of a strict top-down strategy of test use often cannot be tolerated. For these reasons as well as others, researchers and practitioners have often experimented with and implemented other ways of using test scores.

Banding and Cut Scores

One method of reducing the consequences of subgroup differences in test scores and top-down selection is to form bands of test scores that are not considered different, usually using a statistical criterion known as the standard error of the difference, which is based on the reliability of the test. The theory in employment selection use of banding is that the unreliability inherent in most tests makes the people within a band indistinguishable from each other, just as occurs when grades are assigned to students.

Most of us are familiar with a form of banding commonly used in academic situations. Scores on tests are usually grouped into grades (e.g., A, B, C, etc., or red, green, and yellow, as is often the practice in organizational practice) that are reported without specific test score information. So persons with scores of 99 and 93 might both receive an A in a course, just as two with scores of 88 and 85 would receive a B.

Because minorities tend to score lower on cognitive ability tests, creating these bands of indistinguishable scores helps increase the chances that minority applicants will fall in a top band and be hired. There are two ways in which banding can increase minority hiring. One is to make the bands very wide so that a greater number of minority test scorers will be included in the top bands. Of course, a cynic may correctly point out that a test of zero reliability will include everyone in the top band and that this approach supports the use of tests with low reliability. A second way in which to impact the selection of minority individuals is the manner in which individuals are chosen within a band. The clearest way to increase the selection of minority individuals is to choose these persons first within each band before proceeding to consider other individuals in the band, but this has proven difficult to legally justify in U.S. courts (Campion et al., 2001). Other approaches to selection within a band include random selection or selection on secondary criteria unrelated to subgroup status, but these procedures typically do not affect minority hiring rates in practically significant ways (Sackett & Roth, 1991). A discussion of various issues and debates regarding the appropriateness of banding is contained in an edited volume by Aguinis (2004).

Use of Minimum Cut Scores

An extreme departure from top-down selection occurs when an organization sets a minimum cutoff test score such that individuals above some score are selected, whereas those below that score are rejected. In essence, there are two bands of test scores—those judged to represent a passable level of competence and those representing a failing level of test performance. Perhaps the most common use of cutoff scores is in licensing and credentialing, in which the effort is usually to identify a level of expertise and knowledge of the practice of a profession below which a licensure candidate is likely to bring harm to clients. In organizational settings, a cutoff is often required when selection of personnel is done sequentially over time rather than from among a large number of candidates at a single point in time. In this case, hire/reject decisions are made about individuals, and a pass score is essential.

Use of a single cutoff score will certainly reduce the potential utility inherent in a valid test because it ignores the individual differences in ability above the test score cutoff. A great deal of evidence (e.g., Coward & Sackett, 1990) shows that test score–job performance relationships are linear throughout the range of test scores. However, using a minimum cutoff score on a cognitive ability test on which we usually see the largest minority-majority differences to select employees and selecting above that cutoff on a random basis or on the basis of some other valid procedure that does not display subgroup differences (e.g., a personality test where adverse impact is much less likely) may reduce the adverse impact that usually occurs with top-down selection using only a cognitive ability test.

Perhaps the biggest problem with the use of cutoff scores is deriving a justifiable cutoff score. Setting a cutoff is always judgmental. Livingston (1980) and Cascio, Alexander, and Barrett (1988) among others have usually specified the following as important considerations in setting cutoffs: the qualifications of the experts who set the cutoff, the purpose for which the test is being used, and the consideration of the various types of decision errors that can be made (i.e., denying a qualified person and accepting an unqualified individual). One frequently used approach is the so-called Angoff method (Angoff, 1971), in which a representative sample of experts examines each test item and determines the probability that a minimally competent person (the definition and experts' understanding of minimally competent is critical) would answer the question correctly. These probabilities are summed across experts and across items. The result is the cutoff score. A second approach to the setting of cutoff scores is to set them by reference to some acceptable level of performance on a criterion variable. In this case, one could end up saying that an individual with a score of 75 on some test has a 10% (or any percent) chance of achieving success on some job. However, this "benchmarking" of scores against criteria does not resolve the problem because someone will be asked to make sometimes equally difficult decisions about what constitutes acceptable performance. Cizek (2001) provided a comprehensive treatment of methods of setting performance standards.

The use of cutoff scores to establish minimum qualifications or competency is common in licensing exams. Credentialing exams may require evidence of a higher level of skill or performance capability in some domain, but they too usually require only a "pass-fail" decision. Validation of these cutoffs almost always relies solely on the judgments of experts in the performance area of interest. In these cases, careful explication of the behaviors required to perform a set of tasks and the level of "acceptable" performance is essential and likely the only possible form of validation.

Using Profiles of Scores

Another possibility when scores on multiple measures of different constructs are available is that a profile of measured KSAOs is constructed, and this profile is matched to a profile of the KSAOs thought to be required in a job. In this instance, we might measure and quantify the type of job experiences possessed by a job candidate along with their scores on various personality tests, and their oral communications and social skills as measured in an interview and scores on ability tests. If this profile of scores matches that required in the job, then the

person would be selected. This contrasts with the traditional approach described in textbooks in which the person's scores on these tests would be linearly related to performance and combined using a regression model so that each score was optimally linearly related to job performance. In using profiles, one is interested in patterns of scores rather than an optimally weighted composite. Use of profiles of scores presents various complex measurement and statistical problems of which the user should be aware (Edwards, 2002). Instances in which selection decisions are made in this fashion include individual assessments (Jeanneret & Silzer, 1998), which involve the use of multiple techniques using multiple methods of assessment and a clinical judgment by the assessor that a person is qualified for some position (Ryan & Sackett, 1987, 1992, 1998). Another venue in which profiles of test scores are considered is in assessment centers in which candidates for positions (usually managerial) are evaluated in various exercises on different constructs and assessors make overall judgments that are then used in decision making. Overall judgments based on these procedures have shown criterion-related validity (see Ryan & Sackett [1998] for a summary of data relevant to individual assessment and Gaugler, Rosenthal, Thornton, and Bentson [1987] or Arthur, Day, McNelly, and Edens [2003] on assessment center validity), but we are aware of no evidence that validates a profile or configural use of scores. Recently, Davison, Davenport, Yu-Feng, Kory, and Shiyang (2015) have provided a method of estimating the validity of the use of subscores in a profile of scores that may have value in this context.

Perhaps the best description of the research results on the use of profiles to make high-stakes decisions is that we know very little. The following would be some of the issues that should receive research attention: (a) Is a profile of scores actually used, implicitly or explicitly, in combining information about job applicants and what is it? (b) What is the validity of such use and its incremental validity over the use of individual components of the profile or linear composites of the scores in the profile? and (c) What is the adverse impact on various subgroups using profile judgments? Or should a person with a profile above that of another person across the reported scores be selected in favor of a person whose scores are near exact replicas of the desired profile?

Clinical Versus Statistical Judgment

Clinical judgment refers to the use and combination of different types of information to make a decision or recommendation about some person. In psychology, clinical judgment may be most often discussed in terms of diagnoses regarding clinical patients (Meehl, 1954). These judgments are likely quite similar to those made in the individual assessments often used in the selection of high-level executives but also may occur when judgments are made about job applicants in employment interviews, assessment centers, and various other instances in which human resource specialists or psychologists make employment decisions. Clinical judgment is often compared with statistical judgment in which test scores are combined on the basis of an arithmetic formula that reflects the desired weighting of each element of information. The weights may be determined rationally by a group of job experts or by using weights derived from a regression of a measure of overall job success on scores on various dimensions using different methods of measurement. Meehl's original research (1954) showed that the accuracy of the worst regression estimate was equal to the judgments made by human decision makers. A more recent treatment and review of this literature by Hastie and Dawes (2001) has reaffirmed the general conclusion that predictions made by human experts are inferior to those based on a linear regression model. However, human experts are required to identify the types of information used in the prediction task. The predictions themselves are likely best left to some mechanical combination rule if one is interested in maximizing a performance outcome. The overall clinical judgment when used to make decisions should be the focus of the validation effort, but unless it is clear how information is combined by the decision maker, it is unclear what constructs are playing a role in their decisions. The fact that these clinical judgments are often not as highly correlated with externally relevant and important outcomes suggests that at least some of the constructs these decision makers use are not relevant.

In clinical judgment, the presence or absence of adverse impact can be the result of a combination of information that does not display sizable subgroup differences or a bias on the part of the person making the judgment. Psychologists making clinical judgments may mentally adjust scores on the basis of their knowledge of subgroup differences on various measures. There are again no studies of which we are aware that address the use or appropriateness of such adjustments.

SCIENTIFIC OR LONG-TERM PERSPECTIVE: LIMITATIONS OF EXISTING PRIMARY VALIDATION STUDIES

There are a great many meta-analyses of the criterion-related validity of various constructs in the prediction of job performance and many thousands of primary studies. Secondary analyses of meta-analyses have also been undertaken (e.g., Schmidt & Hunter, 1998). The studies that provided these data were nearly all conducted more than 30 years ago. Although it is not necessarily the case that the relationships between ability and performance documented in these studies have changed in the last half-century or so, this database has some limitations. In this section, we describe these limitations and make the case that researchers continue their efforts to evaluate test-performance relationships and improve the quality of the data that are collected.

Concurrent Validation Designs

In criterion-related validation research, concurrent validation studies in which predictor and criterion data are simultaneously collected from job incumbents are distinguished from predictive designs. In the latter, predictor data are collected before hiring from job applicants and criterion data are collected from those hired presumably on the basis of criteria that are uncorrelated with the predictor data after some appropriate period of time when job performance is thought to have stabilized. Defects in the concurrent design (i.e., restriction of range and a different motivational set on the part of incumbents versus applicants) have been described frequently (Barrett, Phillips, & Alexander, 1981). However, some tests are probably more susceptible to motivational differences among job incumbents and applicants, as might be the case for many noncognitive measures that would display differences in validity when the participants in the research were actually being evaluated for employment versus a situation in which they were responding "for research purposes." To our knowledge, this comparison has not been made frequently, and, when it has been done in meta-analyses, cognitive and noncognitive test validities have not been separated (Schmitt, Gooding, Noe, & Kirsch, 1984). Practical considerations have made the use of concurrent designs much more frequent than that of predictive designs (Schmitt et al., 1984).

Meta-analytic data suggest that there are not large differences in the validity coefficients resulting from these two designs. Further, range restriction corrections can be applied to correct for the fact that data for lower-scoring persons are absent from concurrent studies, but these data are often absent in reports of criterion-related research. Nor can we estimate any effects on test scores that might result from the fact that much more is at stake in a testing situation that may result in employment as opposed to one that is being done for research purposes. Moreover, as Sussman and Robertson (1986) maintained, the manner in which some predictive studies are designed and conducted make them little different than concurrent studies.

Unidimensional Criterion Versus Multidimensional Perspectives

Over the last two decades, the view that job performance is multidimensional has become much more widely accepted by I-O psychologists (Borman & Motowidlo, 1997; Campbell, Gasser, & Oswald, 1996). Early validation researchers often used a single rating of what is now called task performance as a criterion, or they combined a set of ratings into an overall performance measure.

In many cases a measure of training success was used as the criterion. The Project A research showed that performance comprised clearly identifiable dimensions (Campbell et al., 1990), and subsequent research has very often included the use of measures of contextual (e.g., helping others) and task performance (Motowidlo, 2003). Some researchers also argue that the nature of what constitutes performance has changed because jobs have changed (Ilgen & Pulakos, 1999). In all cases, the underlying performance constructs should be specified as carefully as possible, perhaps particularly so when performance includes contextual dimensions, which, as is true of any developing literature, have included everything that does not include "core" aspects of a job. Validation studies (and meta-analyses) that include this multidimensional view of performance are very likely to yield information that updates earlier validation results.

Small Sample Sizes

The limitations of small sample sizes in validity research have become painfully obvious with the development of meta-analyses and validity generalization research (Schmidt & Hunter, 1977), as well as the recognition that the power to reject a null hypothesis that there is no test score–performance relationship is very low in much early validation work (Schmidt, Hunter, & Urry, 1976). Although methods to correct for the variability in observed validity coefficients are available and routinely used in meta-analytic and validity generalization research, the use of small samples does not provide for confidence in the results of that research and can be misleading in the short term as enough small sample studies are conducted and reported to discern generalizable findings. This may not be a problem if we are satisfied that the relationships studied in the past are the only ones in which our field is interested, but it is a problem when we want to evaluate new performance models (e.g., models that include a distinction between task, contextual dimensions, or others), new predictor constructs (e.g., some noncognitive constructs or even spatial or perceptual measures), or when we want to assess cross- or multilevel hypotheses. As stated earlier, meta-analyses are not possible in the absence of primary studies.

Inadequate Data Reporting

The impact of some well-known deficiencies in primary validation studies is well known. Corrections for range restriction and criterion unreliability (in the mean and variance of validity coefficients) and for the variability due to small sample size are also well known and routinely applied in validity generalization work. However, most primary studies do not report information that allows for sample-based corrections for criterion unreliability or range restriction. Schmidt and Hunter (1977), in their original meta-analytic effort, used estimates of the sample size of the validity coefficients they aggregated because not even sample size was available in early reports. Consequently, in estimating population validity coefficients, meta-analysts have been forced to use assumed artifact distributions based on the small amount of data that are available. There is some evidence that these assumptions are approximately correct (e.g., Alexander, Carson, Alliger, & Cronshaw, 1989; Sackett & Ostgaard, 1994) for range restriction corrections, but the use of such assumed artifact distributions would not be necessary with adequate reporting of primary data. Unfortunately, such information for most of our primary database is lost. In addition, researchers disagree regarding the appropriate operationalization of criterion reliability (Murphy & DeShon, 2000; Schmidt, Viswesvaran, & Ones, 2000).

HARKing

In psychology, generally, and in organizational psychology (Bosco, Aguinis, Field, & Pierce, in press), there has been increasing concern (Kerr, 1998) about a practice known as HARKing (hypothesizing after the results are known). The implicit hypothesis in criterion-related research

is that the set of predictor variables considered is related to some relevant employee outcome. Not an infrequent practice in this research is to measure a wide range of predictor variables and then report and use only the subset of those predictors that display a statistically (or practically) significant relationship to the outcome variable(s). This is especially problematic when sample sizes are small and there is no cross-validation. The result is likely an overestimate of the validity of the remaining predictors, even when formula estimates of cross-validity are used to estimate shrinkage (Schmitt & Ployhart, 1999).

A related problem is referred to as the file drawer problem in which studies that reveal non-significant correlations are never published or publicly noted. Meta-analysts of criterion-related research will retrieve only those studies available, and if only those that produce significant results are available, then the end result will be an overestimate of the validity of predictors. The potential for this type of bias in the estimation of validity coefficients has been noted by many in selection research (e.g., McDaniel, Rothstein, & Whetzel, 2006). The estimation of the potential for bias in meta-analytic estimates of relationships and the appropriate adjustment of such estimates are available and should be applied to meta-analysis reports (Rothstein, Sutton, & Borenstein, 2005).

Consideration of Multilevel Issues

As described in the section above on the utility and adverse impact associated with selection procedures, selection researchers have attempted to estimate the organizational outcomes associated with the use of valid tests (Boudreau & Ramstad, 2003). Utility is linearly related to validity minus the cost of recruiting and assessing personnel. When multiplied by the number of people and the standard deviation of performance in dollar terms, the estimates of utility for most selection instruments are very large (e.g., see Schmidt, Hunter, Outerbridge, & Trattner, 1986).

Another body of research has focused on the relationship between organizational human resource practices, such as the use of tests and measures of organizational success. The organizational-level research has documented the usefulness of various human resource practices including test use. Terpstra and Rozell (1993) early-reported correlational data that supported the conclusion that organizations that used various selection procedures such as interviews, cognitive ability tests, and biodata had higher annual levels of profit, growth in profit, and overall performance; subsequent research has supported this conclusion (Jiang, Lepak, Hu, & Baer, 2012).

Various other authors have called for multilevel (individuals, work groups, organizations) or cross-level research on the relationship between KSAOs and organizational differences (Schneider, Smith, & Sipe, 2000). Ployhart and Schmitt (2007) and Schneider et al. (2000) have proposed a series of multilevel questions that include considerations of the relationships between the variance of KSAOs and measures of group and organizational effectiveness. In the context of the attraction-selection-attrition model (Schneider, 1987), there are many issues of a multilevel and longitudinal nature that researchers are only beginning to address and about which we have very little or no data. These questions should be addressed if we are to fully understand the relationships between KSAOs and individual and organizational performance. Chapter 5 in this volume provides additional discussion of these issues and questions.

Validation and Long-Term or Scientific Perspective

Given the various limitations of our primary database noted in the previous sections of this chapter, we believe selection researchers should aim to conduct additional large-scale or consortium studies like Project A (Campbell, 1990; Campbell & Knapp, 2001). These studies should include the following characteristics:

1. They should be predictive (i.e., longitudinal with data collection at multiple points), concurrent, and of sufficient sample size to allow for adequate power in the tests of hypotheses. Large-scale studies in which organizations continue data collection over time on an ever-expanding group of participants should be initiated.

2. Multiple criteria should be collected to allow for evaluation of various KSAO–performance relationships.
3. Data should be collected to allow for artifact corrections such as unreliability in the criteria and range restriction.
4. Unit-level data should be collected to allow for evaluation of multilevel hypotheses. These data should include basic unit characteristics and outcome data.
5. Demographic data should be collected to allow for evaluation of subgroup differences in the level of performance and differences in KSAO–performance relationships across subgroups.
6. Data on constructs thought to be related (and unrelated) to the target constructs of interest should be collected to allow for evaluation of broader construct validity issues.
7. Such large-scale studies should include studies of new tests and testing technologies when these become available to allow for innovation.

Obviously, these studies would necessitate a level of cooperation and planning not characteristic of multiple researchers, much less multiple organizations. However, real advancement in our understanding of individual differences in KSAOs and performance will probably not come from additional small-scale studies or meta-analyses of primary studies that address traditional questions with sample sizes, research designs, and measurement characteristics that are not adequate.

CONCLUSIONS

It is certainly true that meta-analyses have provided our discipline with strong evidence that many of the relationships between individual differences and performance are relatively strong and generalizable. However, many situations where validation is necessary do not lend themselves to validity generalization or the use of meta-analytic databases. As a result, practitioners frequently find themselves in situations where well-designed primary studies are required. A focus on the appropriate designs for these studies is therefore important.

Additionally, without primary studies of the relationships between individual differences and performance, there can be no meta-analyses or related applications of validity generalizability and transportability. The quality and nature of the original studies that are the source of our meta-analytic database determine to a great extent the currency and quality of the conclusions derived from the meta-analyses, statistical corrections notwithstanding.

We argue that the field would be greatly served by large-scale primary studies of the type conducted as part of Project A (see Sackett, 1990, or Campbell & Knapp, 2001). These studies should begin with a clear articulation of the performance and predictor constructs of interest. They should involve the collection of concurrent and predictive data and improve upon research design and reporting issues that have bedeviled meta-analytic efforts for the past three decades. Demographic data should be collected and reported. All data should be collected across multiple organizational units and organizations (and perhaps globally), and data describing the organizational context should be collected and recorded. We know much more about the complexities of organizational behavior, research design, measurement, and individual differences than we did 80–100 years ago, and this should be reflected in how we collect our data and make them available to other professionals. The end result will be even greater progress in our understanding of the relationship between individual differences and work performance.

REFERENCES

Ackerman, P. L. (1989). Within-task intercorrelations of skilled performance: Implications for predicting individual differences? (A comment on Henry & Hulin, 1987). *Journal of Applied Psychology, 74*, 360–364.

Aguinis, H. (Ed.) (2004). *Test score banding in human resource selection: Legal, technical and societal issues*. Westport, CT: Praeger.

Alexander, R. A., Carson, K. P., Alliger, G. M., & Cronshaw, S. F. (1989). Empirical distributions of range restricted SD_x in validity studies. *Journal of Applied Psychology, 74*, 253–258.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1954). Technical recommendations for psychological and diagnostic techniques. *Psychological Bulletin, 51*, 201–238.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125–153.

Arvey, R. D., Nutting, S. M., & Landon, T. E. (1992). Validation strategies for physical ability testing in police and fire settings. *Public Personnel Management, 21*, 301–312.

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

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

Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology, 66*, 1–6.

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*, 478–494.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*, 99–109.

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology, 69,* 709–750.

Boudreau, J. W., & Ramstad, P. M. (2003). Strategic industrial and organizational psychology and the role of utility. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology* (Vol. 12, pp. 193–221). Hoboken, NJ: Wiley.

Brogden, H. E. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. *Educational and Psychological Measurement, 11*, 173–195.

Brogden, H. E. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. *Educational and Psychological Measurement, 19*, 181–190.

Brown, C. W., & Ghiselli, E. E. (1953). Percent increase in proficiency resulting from use of selection devices. *Journal of Applied Psychology, 37*, 341–345.

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

Campbell, J. P. (1990). An overview of the army selection and classification project (Project A). *Personnel Psychology, 43*, 231–240.

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

Campbell, J. P., & Knapp, D. J. (Eds.) (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum.

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

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology, 43*, 313–334.

Campion, M. A., Outtz, J. L., Zedeck, S., Schmidt, F. L., Kehoe, J. F., Murphy, K. R., & Guion, R. M. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology, 54*, 149–185.

Cascio, W. F., Alexander, R. A., & Barrett, G. V. (1988). Setting cut scores: Legal, psychometric, and professional issues and guidelines. *Personnel Psychology, 41*, 1–24.

Cascio, W. F., Valenzi, E. R., & Silbey, V. (1978). Validation and statistical power: Implications for applied research. *Journal of Applied Psychology, 63*, 589–595.

Cizek, G. J. (Ed.) (2001). *Setting performance standards*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand-McNally.

Coward, W. M., & Sackett, P. R. (1990). Linearity of ability-performance relationships: A reconfirmation. *Journal of Applied Psychology, 75*, 297–300.

Cropanzano, R., & Wright, T. A. (2003). Procedural justice and organizational staffing: A tale of two paradigms. *Human Resource Management Review. Special Issue: Fairness and Human Resources Management, 13*, 7–39.

Davis, J. E. (1989). Construct validity in measurement: A pattern matching approach. *Evaluation and Program Planning. Special Issue: Concept Mapping for Evaluation and Planning, 12*, 31–36.

Davison, M. L., Davenport, E. C., Jr., Yu-Feng, C., Kory, V., & Shiyang, S. (2015). Criterion-related validity: Assessing the value of subscores. *Journal of Educational Measurement, 52*, 263–279.

Dunnette, M. D. (1963). A note on the criterion. *Journal of Applied Psychology, 47*, 251–254.

Edwards, J. R. (2002). Alternatives to difference scores: Polynomial regression analysis and response surface methodology. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 350–400). San Francisco: Jossey-Bass.

Fiske, D. W. (1951). Values, theory, and the criterion problem. *Personnel Psychology, 4*, 93–98.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analyses of assessment center validity. *Journal of Applied Psychology, 72*, 493–511.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. New York, NY: Wiley.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26*, 461–478.

Gibson, W. M., & Caplinger, J. A. (2007). Transportation of validation results. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence: The professional practice series* (pp. 29–81). Hoboken, NJ: Wiley.

Goldman, B. M., Gutek, B. A., Stein, J. H., & Lewis, K. (2006). Employment discrimination in organizations: Antecedents and consequences. *Journal of Management, 32*(6), 786–830.

Guion, R. M. (1961). Criterion measurement and personnel judgments. *Personnel Psychology, 14*, 141–149.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world*. Thousand Oaks, CA: Sage.

Henry, R. A., & Hulin, C. L. (1989). Changing validities: Ability-performance relations and utilities. *Journal of Applied Psychology, 54*, 365–367.

Hogan, J., & Quigley, A. M. (1986). Physical standards for employment and the courts. *American Psychologist, 41*, 1193–1217.

Hull, C. L. (1928). *Aptitude testing*. Yonkers, NY: World Book.

Ilgen, D. R., & Pulakos, E. D. (Eds.) (1999). *The changing nature of performance*. San Francisco: Jossey-Bass.

Jeanneret, P. R., & Silzer, R. (Eds.) (1998). *Individual psychological assessment: Predicting behavior in organizational settings*. San Francisco, CA: Jossey-Bass.

Jiang, K., Lepak, D. P., Hu, J., & Baer, J. C. (2012). How does human resource management influence organizational outcomes? A meta-analytic investigation of mediating mechanisms. *Academy of Management Journal, 55*, 1264–1294.

Johnson, J. W., & Carter, G. W. (2010). Validating synthetic validation: Comparing traditional and synthetic validity coefficients. *Personnel Psychology, 63*, 755–795.

Kerr, N. L. (1998), HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217.

Lance, C. L., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology, 89*, 22–35.

Lance, C. L., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, S. (2007). Extending the nomological network of assessment center construct validity: Prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance, 20*, 345–362.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183–1192.

Little, R. J. A., & Rubin, D. R. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Livingston, S. A. (1980). Comments on criterion-referenced testing. *Applied Psychological Measurement, 4*, 575–581.

Maier, M. H. (1988). On the need for quality control in validation research. *Personnel Psychology, 41*, 497–502.

McDaniel, M. A. (2007). Validity generalization as a test validation approach. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence: The professional practice series* (pp. 159–180). Hoboken, NJ: Wiley.

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. (2006). Publication bias: A case study of four test vendor manuals. *Personnel Psychology, 59*, 927–953.

Meehl, R. J. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*, 35–44.
Motowidlo, S. J. (2003). Job performance. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology* (Vol. 12, pp. 39–53). Hoboken, NJ: Wiley.
Murphy, K. R. (1986). When your top choice turns you down: The effects of rejected offers on the utility of selection tests. *Psychological Bulletin, 99*, 133–138.
Murphy, K. R. (2009). Content validation is useful for many things, but validity isn't one of them. *Industrial and Organizational Psychology, 2*, 453–464.
Murphy, K. R., & DeShon, R. P. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873–900.
Newman, D. A., Jacobs, R. R., & Bartram, D. (2007). Choosing the best method for local validity estimation: Relative accuracy of meta-analysis versus a local study versus Bayes-analysis. *Journal of Applied Psychology, 92*, 1394–1413.
Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Cognitive ability in selection decisions. In O. Wilhelm (Ed.), *Handbook of understanding and measuring intelligence* (pp. 431–468). Thousand Oaks, CA: Sage.
Peterson, N. G., Wise, L. L., Arabian, J., & Hoffman, R. G. (Eds.) (2001). Synthetic validation and validity generalization: When empirical validation is not possible. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 411–452). Mahwah, NJ: Lawrence Erlbaum.
Picano, J. J., Williams, T. J., & Roland, R. R. (Eds.) (2006). *Assessment and selection of high-risk operational personnel*. New York, NY: Guilford Press.
Ployhart, R. E., & Schmitt, N. (2007). The attraction-selection-attrition model and staffing: Some multilevel implications. In D. B. Smith (Ed.), *The people make the place: Exploring dynamic linkages between individuals and organizations* (pp. 89–102). Mahwah, NJ: Lawrence Erlbaum.
Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing ng organizations: Contemporary practice and theory*. Mahwah, NJ: Lawrence Erlbaum.
Rothstein, H. R. (1992). Meta-analysis and construct validity. *Human Performance, 5*, 71–80.
Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in Meta-analysis: Prevention, Assessment, and Adjustment*. Chichester, UK: Wiley.
Ryan, A. M., & Sackett, P. R. (1987). A survey of individual assessment practices by I/O psychologists. *Personnel Psychology, 40*, 455–488.
Ryan, A. M., & Sackett, P. R. (1992). Relationships between graduate training, professional affiliation, and individual psychological assessment practices for personnel decisions. *Personnel Psychology, 45*, 363–385.
Ryan, A. M., & Sackett, P. R. (1998). Individual assessment: The research base. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment: Predicting behavior in organizational settings* (pp. 54–87). San Francisco, CA: Jossey-Bass.
Sackett, P. R. (Ed.) (1990). Special issue: Project A: The U.S. army selection and classification project. *Personnel Psychology, 43*, 231–378.
Sackett, P. R., & Ostgaard, D. J. (1994). Job-specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology, 79*, 680–684.
Sackett, P. R., & Roth, L. (1991). A Monte Carlo examination of banding and rank order methods of test score use in personnel selection. *Human Performance, 4*, 279–295.
Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist, 56*, 302–318.
Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49*, 929–954.
Scherbaum, C. A. (2005). Synthetic validity: Past, present, and future. *Personnel Psychology, 58*, 481–515.
Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540.
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*, 262–274.
Schmidt, F. L., & Hunter, J. E. (2003). History, development, evolution, and impact of validity generalization and meta-analysis methods, 1975–2001. In K. R. Murphy (Ed.), *Validity generalization: A critical review. Applied Psychology Series* (pp. 31–65). Mahwah, NJ: Lawrence Erlbaum.
Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Trattner, M. H. (1986). The economic impact of job selection methods on size, productivity, and payroll costs of the federal work force: An empirically based demonstration. *Personnel Psychology, 39*, 1–30.
Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validity studies. *Journal of Applied Psychology, 61*, 473–485.

Schmidt, F. L., Mack, M. J., & Hunter, J. E. (1984). Selection utility in the occupation of U.S. park ranger for three modes of test use. *Journal of Applied Psychology, 69*, 490–497.

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

Schmitt, N., Gooding, R., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982, and the investigation of study characteristics. *Personnel Psychology, 37*, 407–422.

Schmitt, N., & Ployhart, R. E. (1999). Estimates of cross-validity for stepwise regression and with predictor selection. *Journal of Applied Psychology, 84*, 50–57.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82*, 717–730.

Schneider, B. (1987). The people make the place. *Personnel Psychology, 40*, 437–454.

Schneider, B., Smith, D., & Sipe, W. P. (2000). Personnel selection psychology: Multilevel considerations. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 91–120). San Francisco, CA: Jossey-Bass.

Scott, W. D. (1915). The scientific selection of salesmen. *Advertising and Selling, 5*, 5–7.

Shotland, A., Alliger, G. M., & Sales, T. (1998). Face validity in the context of personnel selection: A multimedia approach. *International Journal of Selection and Assessment, 6*, 124–130.

Society for Industrial and Organizational Psychology. (1987). *Principles for the validation and use of personnel selection procedures.* (3rd ed.). College Park, MD: Author.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures.* (4th ed.). Bowling Green, OH: Author.

Strickland, W. J. (1979). The relationship between program evaluation research and selection system validation: Application to the assessment center method. *Dissertation Abstracts International, 40*(1-B), 481–482.

Sussman, M., & Robertson, D. U. (1986). The validity of validity: An analysis of validation study designs. *Journal of Applied Psychology, 71*, 461–468.

Terpstra, D. E., & Kethley, R. B. (2002). Organizations' relative degree of exposure to selection discrimination litigation. *Public Personnel Management, 31*, 277–292.

Terpstra, D. E., & Rozell, E. J. (1993). The relationship of staffing practices to organizational level measures of performance. *Personnel Psychology, 46*, 27–48.

Thayer, P. W. (1992). Construct validation: Do we understand our criteria? *Human Performance, 5*, 97–108.

Thorndike, E. L. (1911). *Individuality.* Boston, MA: Houghton Mifflin.

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

Zyphur, M. J., Oswald, F. L., & Rupp, D. E. (2015). Rendezvous overdue: Bayes analysis meets organizational research. *Journal of Management, 41*, 387–389.