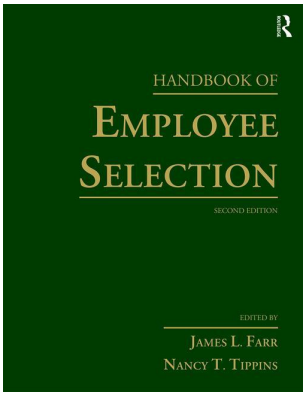

Handbook of Employee Selection

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

Validity Considerations in the Design and Implementation of Selection Systems

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-3
Jerard F. Kehoe, Paul R. Sackett
Published online on: 22 Mar 2017

3

Validity Considerations in the Design and Implementation of Selection Systems

JERARD F. KEHOE AND PAUL R. SACKETT

Validity, along with reliability, is a concept that provides the scientific foundation upon which we construct and evaluate predictor and criterion measures of interest in personnel selection. It offers a common technical language for discussing and evaluating the accuracy of inferences we desire to make based on those scores (e.g., high scores on our predictor measure are associated with high levels of job performance; high scores on our criterion measure are associated with high levels of job performance).[1] Furthermore, the literature surrounding validity provides a framework for scientifically sound measure development that, a priori, can enable us to increase the likelihood that scores resulting from our measures will be generalizable, and inferences we desire to make based upon them, supported.

Like personnel selection itself, science and practice surrounding the concept of validity continue to evolve, with changes affecting not only its evaluation but also its very definition, as evidenced by comparing editions of the *Standards for Educational and Psychological Testing* produced over the past half century by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (AERA, APA, & NCME, 2014). The evolution of validity has been well covered in the personnel selection literature (e.g., Binning & Barrett, 1989; McPhail, 2007; Schmitt & Landy, 1993; Society for Industrial and Organizational Psychology, 2003), and will continue to be well covered in this Handbook. This chapter and the two chapters immediately before and after all speak directly to developments with regard to validity, particularly as it relates to personnel selection. The contribution of this chapter is to develop a more comprehensive understanding of the manner in which the design and implementation of operational selection systems have implications for validity.

OVERVIEW

We begin with a conceptual treatment of validity as it is represented in the personnel selection profession. This treatment attempts to outline a set of distinctions that we view as central to an understanding of validity. Namely, we discuss (a) validity as predictor-criterion relationship versus broader conceptualizations, (b) validity of an inference versus validity of a test, (c) types of validity evidence versus types of validity, (d) validity as an inference about a test score versus

validation as a strategy for establishing job relatedness, (e) the predictive inference versus the evidence for it, and (f) validity limited to inferences about individuals versus including broader consequences of test score use. Our belief is that a clear understanding of these foundational issues in validity is essential for effective research and practice in the selection arena. In addition, we believe that this conceptual foundation should guide the treatment we give below to the operational and practical considerations for establishing that a particular selection system is supported by persuasive validity evidence.

Following this conceptual treatment of validity, we describe key validity considerations in the design and development of operational selection systems. For each of these considerations we describe the manner in which they can strengthen or weaken conclusions about the validity of the selection system as implemented for its intended purpose(s).

PART 1: CONCEPT OF VALIDITY

Validity, according to the 2014 *Standards for Educational and Psychological Testing*, is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11). There is a long history and considerable literature on the subject of validity. With limited space here, it is impossible to do justice to the subject. We attempt to highlight a set of important issues in the ongoing development of thinking about validity, but we direct the interested reader to a set of key resources for a strong foundation on the topic. One key set of references is the set of chapters on the topic of validity in the four editions of *Educational Measurement*, which is that field's analog to the *Handbook of Industrial and Organizational Psychology*. Cureton (1951), Cronbach (1971), Messick (1989), and Kane (2006) each offer detailed treatment of the evolving conceptualizations of validity. Another key set focuses specifically on validity in the context of personnel selection. Two prominent articles on validity in the employment context have been published in the *American Psychologist* by Guion (1974) and Landy (1986). There is also a very influential paper by Binning and Barrett (1989). A third key set is made up of classic highly cited articles in psychology—Cronbach and Meehl's (1955) and Loevinger's (1957) treatises on construct validity.

Our focus in this section is entirely conceptual. This chapter does not address operational issues in the design of research studies aimed at obtaining various types of validity evidence except to the extent that local studies might be conducted within the process of design and development. Rather, we attempt to outline a set of issues that we view as central to an understanding of validity.

Validity as Predictor-Criterion Relationship Versus Broader Conceptualizations

In the first half of the 20th century, validity was commonly viewed solely in terms of the strength of predictor-criterion relationships. Cureton's (1951) chapter on validity stated, reasonably, that validity addresses the question of "how well a test does the job it was employed to do" (p. 621). But the "job it was employed to do" was viewed as one of prediction, leading Cureton to state that, "Validity is . . . defined in terms of the correlation between the actual test scores and the 'true' criterion measures" (pp. 622–623).

But more questions were being asked of tests than whether they predicted a criterion of interest. These included questions about whether mastery of a domain could be inferred from a set of questions sampling that domain and about whether a test could be put forward as a measure of a specified psychological construct. A landmark event in the intellectual history of the concept of validity was the publication of the first edition of what is now known as the *Standards for Educational and Psychological Testing* (APA, 1954), in which a committee headed by Lee Cronbach, with Paul Meehl as a key member, put forward the now familiar notions of predictive, concurrent, content, and construct validity. Cronbach and Meehl (1955) elaborated their position on construct validity a year later in their seminal *Psychological Bulletin* paper. Since then, validity has

been viewed more broadly than predictor-criterion correlations, with the differing validity labels viewed first as types of validity and more recently as different types of validity evidence or as evidence relevant to differing inferences to be drawn from test scores.

Validity of an Inference Versus Validity of a Test

Arguably the single most essential idea regarding validity is that it refers to the degree to which evidence supports the inferences one proposes to draw about the target of assessment (in the I-O world, most commonly an individual; in other settings, a larger aggregate such as a classroom or a school) from their scores on assessment devices. The generic question "Is this a valid test?" is not a useful one; rather, the question is "Can a specified inference about the target of assessment be validly drawn from scores on this device?" Several important notions follow from this position.

First, it thus follows that the inferences to be made must be clearly specified. Multiple inferences are frequently proposed. Consider a technical report stating, "This test representatively samples the established training curriculum for this job. It measures four subdomains of job knowledge, each of which is predictive of subsequent on-the-job task performance." Note that three claims are made here, dealing with sampling, dimensionality, and prediction, respectively. Each claim is linked to one or more inferences about a test taker (i.e., degree of curriculum mastery, differentiation across subdomains, relationships with subsequent performance, and incremental prediction of performance across subdomains).

Second, support for each inference is needed to support the multifaceted set of claims made about inferences that can be drawn from the test. Each inference may require a different type of evidence. The claim of representative content sampling may be supported by evidence of the form historically referred to as "content validity evidence," namely, a systematic documentation of the relationship between test content and job knowledge requirements, typically involving the judgment of subject matter experts. The claim of multidimensionality may be supported by factor-analytic evidence, and evidence in support of this claim is one facet of what has historically been referred to as construct validity evidence (i.e., evidence regarding whether the test measures what it purports to measure). The claim of prediction of subsequent task performance may be supported by what has historically been referred to as "criterion-related validity evidence," namely, evidence of an empirical relationship between test scores and subsequent performance. Note that the above types of evidence are provided as examples; multiple strategies may be selected alone or in combination as the basis for support for a given inference. For example, empirical evidence of a test-criterion relationship may be unfeasible in a given setting because of sample size limitations, and the investigator may turn to the systematic collection of expert judgment as to the likelihood that performance on various test components is linked to higher subsequent job performance.

Third, some proposed inferences receive support as evidence is gathered and evaluated, whereas others do not. In the current example, what might emerge is strong support for the claim of representative sampling and strong support for the claim of prediction of subsequent performance, but evidence of a unidimensional rather than the posited multidimensional structure. In such cases, one should revise the claims made for the test; in this case, dropping the claim that inferences can be drawn about differential standing on subdomains of knowledge.

Types of Validity Evidence Versus Types of Validity

Emerging from the 1954 edition of what is now the *Standards for Educational and Psychological Testing* was the notion of multiple types of validity. The triumvirate of criterion-related validity, content validity, and construct validity came to dominate writings about validity. At one level, this makes perfect sense. Each of these deals with different key inferences one may wish to draw about a test. First, in some settings, such as many educational applications, the key inference is

one of content sampling. Using tests for purposes such as determining whether a student passes a course, progresses to the next grade, or merits a diploma relies heavily on the adequacy with which a test samples the specified curriculum. Second, in some settings, such as the study of personality, the key inference is one of appropriateness of construct labeling and specification. There is a classic distinction (Loevinger, 1957) between two types of construct validity questions, namely, questions about the existence of a construct (e.g., Can one define a construct labeled "integrity" and differentiate it from other constructs?) and questions about the adequacy of a given measure of a construct (e.g., Can test X be viewed as a measure of integrity?). Third, in some settings, such as the personnel selection setting of primary interest for the current volume, the key inference is one of prediction: Can scores from measures gathered before a selection decision be used to draw inferences about future job behavior? Criterion-related validity evidence is a central mechanism for establishing this inference, though content-related evidence also supports this inference, as discussed below.

Over the last several decades, there has been a move from viewing these as types of validity to types of validity evidence. All lines of evidence—content sampling, dimensionality, convergence with other measures, investigations of the processes by which test takers respond to test stimuli, or relations with external criteria—deal with understanding the meaning of test scores and the inferences that can be drawn from them. Because construct validity is the term historically applied to questions of the meaning of test scores, the position emerged that if all forms of validity evidence contributed to understanding the meaning of test scores, then all forms of validity evidence were really construct validity evidence. The 1999 and 2014 editions of the *Standards* pushed this one step further: If all forms of evidence are construct validity evidence, then "validity" and "construct validity" are indistinguishable. Thus, the *Standards* refer to "validity" rather than "construct validity" as the umbrella term. This seems useful, because construct validity carries the traditional connotations of referring to specific forms of validity evidence, namely convergence with conceptually related measures and divergence from conceptually unrelated measures.

Thus, the current perspective reflected in the 2014 *Standards* is that validity refers to the evidentiary basis supporting the inferences that a user claims can be drawn from a test score. Many claims are multifaceted, and thus multiple lines of evidence may be needed to support the claims made for a test. A common misunderstanding of this perspective on validity is that the test user's burden has been increased, because the user now needs to provide each of the types of validity evidence. In fact, there is no requirement that all forms of validity evidence be provided; rather, the central notion is, as noted earlier, that evidence needs to be provided for the inferences that one claims can be drawn from test scores. For example, if one's intended inferences make no claims about content sampling, then content-related evidence is not needed. If the claim is simply that scores on a measure can be used to forecast whether an individual will voluntarily leave the organization within a year of hire, then the only inference that needs to be supported is the predictive one. One may rightly assert that scientific understanding is aided by obtaining other types of evidence than those drawn on to support the predictive inference (i.e., forms of evidence that shed light on the construct(s) underlying test scores), but we view such evidence gathering as desirable but not essential. One's obligation is simply to provide evidence in support of the inferences one wishes to draw.

Validity as an Inference About a Test Score Versus Validation as a Strategy for Establishing Job Relatedness

In employment settings, the most crucial inference to be supported about any measure is whether the measure is job-related. Labeling a measure as job-related means "scores on this measure can be used to draw inferences about an individual's future job behavior"—we term this the "predictive inference." In personnel selection settings, our task is to develop a body of evidence to support the predictive inference. The next section of this chapter outlines mechanisms for doing so.

Some potential confusion arises from the failure to differentiate between settings where types of validity evidence are being used to draw inferences about the meaning of test scores rather than to draw a predictive inference. For example, content-related validity evidence refers to the adequacy with which the content of a given measure samples a specified content domain. Assume that one is attempting to develop a self-report measure of conscientiousness to reflect a particular theory that specifies that conscientiousness has four equally important subfacets: dependability, achievement striving, dutifulness, and orderliness. Assume that a group of expert judges is given the task of sorting the 40 test items into these four subfacets. A finding that 10 items were rated as reflecting each of the four facets would support the inference of adequate domain sampling and contribute to an inference about score meaning. Note that this inference is independent of the question about the job relatedness of this measure. One could draw on multiple lines of evidence to further develop the case for this measure as an effective way to measure conscientiousness (e.g., convergence with other measures) without ever addressing the question of whether predictive inferences can be drawn from this measure for a given job. When one's interest is in the predictive hypothesis, various types of validity evidence can be drawn upon to support this evidence, as outlined below.

Predictive Inference Versus the Evidence for It

As noted above, the key inference in personnel selection settings is a predictive one, namely the inferences that scores on the test or other selection procedure can be used to predict the test takers' subsequent job behavior. A common error is the equating of the type of inference to be drawn with the type of evidence needed to support the inference. Put more bluntly, the error is to assert that, "If the inference is predictive, then the needed evidence is criterion-related evidence of the predictive type."

Scholars in the I-O area have clearly articulated that there are multiple routes to providing evidence in support of the predictive hypothesis. Figure 3.1 presents this position in visual form. Models of this sort are laid out in Binning and Barrett (1989) and in the 2014 *Standards*. This upper half of Figure 3.1 shows a measured predictor and a measured criterion. Because both are measured, the relationship between these two can be empirically established. The lower half of Figure 3.1 shows an unmeasured predictor construct domain and an unmeasured criterion construct domain. Of interest are the set of linkages among the four components of this model.

The first and most central point is that the goal of validation research in the personnel selection context is to establish a linkage between the predictor measure (Figure 3.1, upper left) and the criterion construct domain (Figure 3.1, lower right). The criterion construct domain is the conceptual specification of the set of work behaviors that one is interested in predicting. This criterion construct domain may be quite formal and elaborate, as in the case of a job-analytically-specified set of critical job tasks, or it may be quite simple and intuitive, as in the case of an organization that asserts that it wishes to minimize voluntary turnover within the first year of employment and thus specifies this as the criterion domain of interest.

The second central point is that there are three possible mechanisms for linking an observed predictor score and a criterion construct domain. The first is via a sampling strategy. If the predictor measure is a direct sample of the criterion construct domain, then the predictive inference is established based on expert judgment (e.g., obtained via a job analysis process) (Linkage 5 in Figure 3.1). Having an applicant for a symphony orchestra position sight read unfamiliar music is a direct sample of this important job behavior. Having an applicant for a lifeguard position dive to the bottom of a pool to rescue a simulated drowning victim is a simulation, rather than a direct sample of the criterion construct domain. However, it does rely on domain sampling logic and, like most work sample tests, aims at psychological fidelity in representing critical aspects of the construct domain.

The second mechanism for linking an observed predictor and a criterion construct domain is via establishing a pair of linkages, namely (a) the observed predictor–observed criterion link (Linkage 1 in Figure 3.1) and (b) the observed criterion–criterion construct domain link

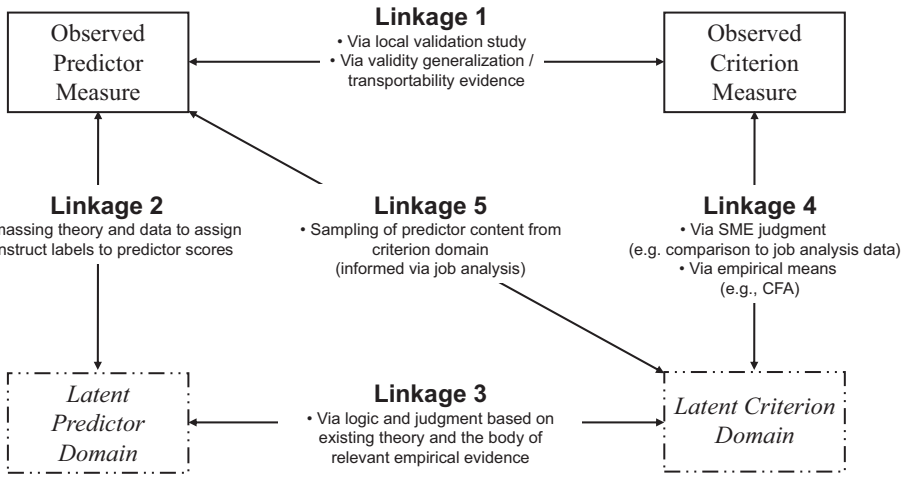

FIGURE 3.1 Routes to Providing Evidence in Support of the Predictive Inference

(Adapted from Binning, J. F., & Barrett, G. V., *Journal of Applied Psychology*, *74*, 478–494, 1989.)

(Linkage 4 in Figure 3.1). The first of these links can be established empirically, as in the case of local criterion-related evidence, or generalized or transported evidence. Critically, such evidence must be paired with evidence that the criterion measure (e.g., ratings of job performance) can be linked to the criterion construct domain (e.g., actual performance behaviors). Such evidence can be judgmental (e.g., comparing criterion measure content to critical elements of the criterion construct domain revealed through job analyses) and empirical (e.g., fitting CFA models to assess whether dimensionality of the observed criterion scores is consistent with the hypothesized dimensionality of the criterion construct domain). It commonly involves showing that the chosen criterion measures do reflect important elements of the criterion construct domain. Observed measures may fail this test, as in the case of a classroom instructor who grades solely on attendance when the criterion construct domain is specified in terms of knowledge acquisition, or in the case of a criterion measure for which variance is largely determined by features of the situation rather than by features under the control of the individuals.

The third mechanism also focuses on a pair of linkages, namely (a) linking the observed predictor scores and the predictor construct domain (Linkage 2 in Figure 3.1) and (b) linking the predictor construct domain and the criterion construct domain (Linkage 3 in Figure 3.1). The first linkage involves obtaining data to support interpreting variance in predictor scores as reflecting variance in a specific predictor construct domain. This reflects one form of what has historically been referred to as construct validity evidence, namely, amassing theory and data to support assigning a specified construct label to test scores. For example, if a test purports to measure achievement striving, one might offer a conceptual mapping of test content and one's specification of the domain of achievement striving, paired with evidence of empirical convergence with other similarly specified measures of the construct. However, showing that the measure does reflect the construct domain is supportive of the predictive inference only if the predictor construct domain can be linked to the criterion construct domain. Such evidence is logical and judgmental, requiring a clear articulation of the basis for asserting that individuals who are higher in the domain of achievement striving will have higher standing on the criterion construct domain than individuals lower in achievement striving.

Thus, there are multiple routes to establishing the predictive inference. These are not mutually exclusive; one may provide more than one line of evidence in support of the predictive inference. The type of measure does not dictate the type of evidentiary strategy chosen.

Validity Limited to Inferences About Individuals Versus Including Broader Consequences of Test Score Use

In the last two decades, considerable attention has been paid to new views of validity that extend beyond the inferences that can be drawn about individuals to include a consideration of the consequences of test use. The key proponent of this position is Messick (1989). Messick noted that it is commonly asserted that the single most important attribute of a measure is its validity for its intended uses. He noted that at times test use has unintended negative consequences, as in the case in which a teacher abandons many key elements of a curriculum to focus all effort on preparing students to be tested in one subject. Even if inferences about student domain mastery in that subject can be drawn with high accuracy, Messick argued that the negative consequences (i.e., ignoring other subjects) may be so severe as to argue against the use of this test. If validity is the most important attribute of a test, then the only way for negative consequences to have the potential to outweigh validity evidence in a decision about the appropriateness of test use was for consequences of test use to be included as a facet of validity. Thus, he argued for a consideration of traditional aspects of validity (which he labeled "evidential") and these new aspects of validity (which he labeled "consequential"). These ideas were generally well received in educational circles, and the term "consequential validity" came to be used; these ideas, however, were not well received in the I-O field. In this usage, a measure with unintended negative consequences lacks consequential validity. This perspective views such negative consequences as invalidating test use.

The 2014 *Standards* rejects this view. Although evidence of negative consequences may influence decisions concerning the use of predictors, such evidence will only be related to inferences about validity if the negative consequences can be directly traced to the measurement properties of the predictor. Using an example that one of us (Sackett) contributed to the SIOP *Principles for the Validation and Use of Personnel Selection Procedures* (2003), consider an organization that (a) introduces an integrity test to screen applicants, (b) assumes that this selection procedure provides an adequate safeguard against employee theft, and (c) discontinues use of other theft-deterrent methods (e.g., video surveillance). In such an instance, employee theft might actually increase after the integrity test is introduced and other organizational procedures are eliminated. Thus, the intervention may have had an unanticipated negative consequence on the organization. These negative consequences do not threaten the validity of inferences that can be drawn from scores on the integrity test, because the consequences are not a function of the test itself.

Given recent encouragement to evaluate selection systems with respect to their influence on organization-level outcomes (e.g., Ployhart & Weekley, Chapter 5 in this volume), we think it is helpful to distinguish between the validity of test scores used to select individuals into organizations and an evaluation of aggregate consequences of the selection system at the organization level. Consider the example of a measure of service orientation used to hire customer service employees for the purpose of improving customer satisfaction. Our view is that even where there is evidence that customer service employees selected for service orientation improve organization-level results (e.g., higher average district-level customer retention), such organization-level outcomes do not constitute evidence of validity for service orientation scores. To be sure, such evidence of organization-level impact would likely have great importance to the organization and may even be far more important to the organization than evidence showing that employees with higher service orientation produce higher customer satisfaction, but such evidence would not strengthen the validity claim that employees with higher service orientation scores produce more satisfied customers. In the design of this selection system, the I-O psychologist relies on information about the conceptual linkage at the *individual level* between employee attributes and employee outcomes as the basis for choosing and validating an assessment measuring service orientation.

This is a nuanced but important point for this chapter. Our focus in this chapter is on the validity of inferences about selection procedure scores that operate at the level of individual applicants and employees. In effect, our point is that for validity inferences to be meaningful both theoretically and practically, the predictor scores and the intended outcome results (criteria) must be at the same level of analysis. This does not mean that validity claims cannot be made with respect to organization-level purposes. In this example, one could evaluate a claim that an

organization-level measure of customer service, such as a district-average rating of customer satisfaction, is a valid predictor of district-level customer retention.

We should say that successful validation at the individual level is not the ultimate objective for the I-O professional who is designing and implementing a selection system. The most important objective is that the selection system be useful and lead to the outcomes intended by the organization and the people in the organization. While valid selection procedures may be one of the most important contributions the I-O psychologist can make to organization-level outcomes, validity is not, in general, sufficient to ensure organization-level results. Indeed, the selection professional is very likely to have a primary focus on making design and implementation decisions that maximize the intended outcomes at both individual and organizational levels, confident in the science-based conclusion that validity is an important building block for producing an effective selection system.

Summary of Part 1

In conclusion, we have attempted to develop six major points about validity. These are that (1) we have moved far beyond early conceptualizations of validity as the correlation between test scores and criterion measures; (2) validity is not a characteristic of a test, but rather refers to inferences made from test scores; (3) we have moved from conceptualizing different types of validity to a perspective that there are different types of validity evidence, all of which contribute to an understanding of the meaning of test scores; (4) the key inference to be supported in employment settings is the predictive inference, namely, that inferences about future job behavior can be drawn from test scores; (5) there are multiple routes to gathering evidence to support the predictive inferences; and (6) although evidence about unintended negative consequences (or intended positive consequences) of test use (e.g., negative applicant reactions to the test) may affect a policy decision as to whether or not to use the test, such evidence is not a threat to the predictive inference and does not affect judgments about the validity of the test. Our belief is that a clear understanding of these foundational issues in validity is essential for effective research and practice in the selection arena.

PART 2: VALIDITY CONSIDERATIONS IN THE DESIGN AND IMPLEMENTATION OF SELECTION SYSTEMS

The following discussion of validity considerations in the design and implementation of selection systems begins with our perspective about the meaning of validity with regard to selection systems and follows by describing how design and development decisions in each of six major stages of a prototypic design and development process can generate evidence for three types of inferences supporting a conclusion of validity. These stages are (1) specify the intended uses and outcomes of the selection process; (2) describe the work; (3) choose/develop predictor and criterion assessment processes; (4) prescribe the manner in which assessment scores are to be used; (5) prescribe the policies and rules that govern the operation of the selection system; and (6) manage and maintain the selection system over time. These stages are roughly sequential but may overlap considerably.

The three types of inferences are about (1) the intended uses and outcomes, (2) the quality of predictor and criterion scores, and (3) the prediction rationale.

The Meaning of Validity in the Context of Selection Systems

In the domain of personnel selection, the *Standards* and *Principles* distinguish the definition of validity from other types of inferences relating to intended outcomes where those outcomes (consequences) "do not follow directly from test score interpretations" (*Standards*, p. 21; related

comments in *Principles* on p. 7). However, just as the *Standards* describes a professional responsibility to provide evidence of validity, it also describes a professional responsibility to support claims about outcomes that do not follow from test score interpretations. We adhere to this distinction in this chapter about validity considerations. The subsections below address the design and implementation considerations relating to evidence of validity and only briefly acknowledge certain key efficacy considerations, for example, relating to cut scores.

Note, throughout this discussion we distinguish between uses and outcomes of selection processes. "Uses" refers to the particular human resources (HR) process supported by selection including, for example, external hiring, internal lateral movement, internal progression programs, promotion/demotion decisions, selection into training/development programs, and downsizing. Different HR processes may have somewhat different implications for the design, implementation, and validity evidence of the supporting selection system. In contrast, the language of "outcomes" is used here to refer to the specific, individual-level work behaviors, products, tasks, contextual behaviors, etc. that the organization intends to influence with the selection system. The *Standards* refers to "proposed uses" in a very broad sense to include both uses and outcomes. We make this distinction because intended uses and intended outcomes can have different consequences for design, implementation, and validation of the selection system.

A Framework for Describing Validity Considerations in the Design and Implementation of Selection Systems

We propose a two-dimensional framework to organize information about the manner in which design and implementation decisions influence inferences about the local validity of selection test scores. The two dimensions of this framework are (1) type of inference and (2) stage of the design and implementation process. In this chapter we describe the manner in which evidence for each of three types of inference is gathered (or not) in each of the six stages of design and implementation. Table 3.1 displays this framework.

TABLE 3.1

A Framework for Describing Validity Considerations in the Design and Implementation of Selection Systems

	Key Inferences in the Selection Validity Rationale		
Stages of Design and Implementation	*Intended Uses and Outcomes*	*Quality of Predictor and Criterion Scores*	*Prediction Rationale*
Stage 1: Specify Intended Uses and Outcomes	• Identify types of selection processes (uses) • Identify most important intended outcomes o Usually it is not feasible to address all desirable outcomes • Evidence that outcomes are a function of individual differences • Determine that no artificial barriers block a predictive relationship • Confirm scope of intended outcomes • Ensure sufficient authority to specify intentions	None from this stage of work	No direct evidence for any prediction rationale is generated at this stage.

	Key Inferences in the Selection Validity Rationale		
Stages of Design and Implementation	*Intended Uses and Outcomes*	*Quality of Predictor and Criterion Scores*	*Prediction Rationale*
Stage 2: Describe the Work	• Sufficient scope and authority to describe the meaning and relevance of intended outcomes and associated work behaviors • Identify important work elements and worker behavior associated with intended outcomes • Provide expertise required for credible work information	• Provide necessary expertise to provide credible information needed to inform predictor constructs as requirements for successful outcomes/ work behaviors • Distinguish between three types of relevant expertise o Work content o Importance to work o Assessment-related expert judgment	Requires sufficient expertise to: • Establish rational/ observed link between worker attributes (predictor constructs) and work outcomes/ behaviors • Provide credible information required by synthetic validation procedures
Stage 3: Choose/ Develop Predictor and Criterion Assessment Processes	• Choice/development directed by information generated in Stages 1 and 2 • Both validity and usefulness influence the choice or development of predictors	• Online, unproctored predictor assessment • Choice of personality scale scores to suit specifics of local outcomes • Absence of bias in criterion measures	• Importance of alignment between predictor and criterion constructs • Implications of complex, multidimensional outcomes • Choosing among commercially available predictors • Incremental contributions to prediction of complex criteria • Expert judgment required to generalize validity conclusions from previous research
Stage 4: Prescribe Score Usage	• No implications for validity evidence relating to intended uses or outcomes	• Using predictor scores to inform selection decision makers' judgments	• Creating composite scores from weighted predictor measures • Predictor scores used in sequence
Stage 5: Prescribe Governing Policies/ Rules	• No implications because equivalencies, exemptions, and waivers are based on considerations unrelated to any interpretation of test scores	• Test-taking conditions re: retesting, test preparation, and online administration are likely to affect scores and alter the generalizability of past validity research to the local setting	• Changes to test-taking conditions designed to accommodate disabilities will have a largely unknown impact on the generalizability of previous research about prediction properties to local setting
Stage 6: Manage and Maintain the Selection System	Adapting to dynamic validity factors • Monitor the importance of intended uses and outcomes • Track metrics for achieved outcomes • Link outcomes to selection results (not validation)	• Test administration and scoring processes may alter the generalizability of previous research about predictor measurement properties to local setting • Train selection administration staff re: process standards • Track predictor score characteristics • Audit threats to predictor score meaning and quality	• Maintain the current professional expertise of the personnel selection expert

Three Types of Key Inferences

In this framework, we describe three types of key inferences as being critical to claims of validity for selection tests. These are (1) inferences relating to the intended uses and outcomes themselves, (2) inferences related to the psychometric quality of predictor and criterion scores including properties such as reliability, group differences, construct validity, and measurement bias, and (3) inferences related to the predictive relationship between test scores and important outcomes.

Intended Uses and Outcomes The first category of inferences relates to attributes of the intended uses and outcomes associated with the purposes of the selection system. Certainly, the specification of intended uses and outcomes is necessary to construct an appropriate validation process. However, it is also necessary that the designer makes certain inferences about the intended uses and outcomes in order to design valid selection procedures. For example, intended outcomes will be amenable to a selection solution only if the designer can infer that, in the local context, they are a function of stable individual differences to some meaningful degree.

Quality of Predictor and Criterion Scores Inferences about the quality of predictor and, where feasible and needed, criterion scores are central to any evaluation of the validity of the predictor scores. This is a category of diverse inferences, all of which relate to some aspect of the psychometric quality of the predictor and criterion scores, as used, including that (a) the selected predictor and criterion assessment methods measure the intended predictor and criterion constructs; (b) the predictor and criteria scores generated in the local setting are reliable; (c) the local assessment processes themselves do not introduce unique sources of measurement error or bias; (d) the manner in which predictor scores are used does not change the meaning of the scores; and (e) the local measurement and usage conditions are supported by relevant previous research about the psychometric qualities of the predictor measures.

This type of inference applies to all measured predictors to which the validity inference applies including resume-based accomplishment/qualification algorithms, interviews, structured skill/ability tests, psychological inventories, work samples, job knowledge and situational judgment tests, structured exercises, and manager/HR staff ratings and judgments. However, as a practical matter, some predictable outcomes of the selection process (possible criteria) are almost always not specified as criteria for a validation effort. This point simply acknowledges that a comprehensive identification and assessment of all desirable selection outcomes is usually not feasible. This is because (a) the list of all desirable outcomes is very long in many cases and some are more important or salient than others, and (b) it is not feasible to assess certain desirable outcomes due to constraints in time, cost, resources, or assessment methodology. For example, organizations are likely to always value the benefits in safety, employee health, and avoided human and dollar costs that are predictably a result of the use of cognitive ability predictors in selection systems, but these desirable outcomes are frequently not salient to the purposes of the selection system and/or may not be assessable by any feasible process.

Prediction Rationale The prediction rationale refers to the evidence and reasoning supporting the claim that scores on the predictor, as used, are predictive of the intended outcomes. This type of inference is central to the claim that selection test scores are valid. The necessary but insufficient foundation for this rationale is that the chosen predictor constructs are conceptually linked to the important intended outcomes (criterion constructs) in a manner that implies predictor constructs will be predictive of criterion constructs. The discussion in the first part of this chapter provides a detailed analysis of the meaning of this category of inferences and distinguishes its meaning from the variety of types of evidence that can be used to inform this inference.

Six Stages of Design and Implementation

We organize the design and implementation work into six major stages to help describe the manner in which design and implementation can provide evidence relevant to the three types of validity inferences. Although these stages are in a roughly logical order, they are interdependent and may significantly overlap with one another. These stages can produce different evidence and support different inferences about the validity of the predictor scores in the system. These six stages are (1) specifying the intended uses and outcomes, (2) describing the work, (3) developing (or choosing) the predictor and criterion assessment procedures, (4) determining the manner in which the predictor scores will be used to make the personnel decisions, (5) establishing the policies and rules that will govern the operation of the selection system, and (6) managing and maintaining the selection system.

To be sure, these six stages represent a prototypic model of design and implementation work and can vary greatly across circumstances. Nevertheless, while any of these may or may not be a stage of actual design and implementation work, each is associated with a set of considerations critical to any selection system. For example, even in the implausible case where no design or implementation work addressed the manner in which test scores should be used, test scores will be used in practice in some fashion. For this reason, each of the six stages represent considerations that have implications for the validity of every selection system regardless of the amount of attention and expertise, if any, that was invested during development.

Throughout this treatment, we presume that the design and implementation work is carried out by professionals with personnel selection expertise. There's little point in describing the prototypic work of non-experts.

For each of these six stages, the subsections below describe the types of validity evidence generated or relied upon in that stage and the types of validity-related inferences supported by that evidence.

Stage 1: Specify Intended Uses and Outcomes

The primary role of this stage of activities is to specify the intended uses and outcomes, which provide the direction needed to begin designing the system including the choice of predictor constructs and measures, the purpose and types of work analyses, and the consideration of most appropriate and feasible types of evidence supporting the predictive relationship between the selection procedures and outcomes.

Inferences About Intended Uses and Outcomes

Generally, the specification of uses and outcomes does not, itself, generate evidence of validity. Rather, it specifies the intended uses and outcomes that will define the criteria for the predictive inference and suggest the types of predictors likely to influence the target criteria. Nevertheless, the specified uses and outcomes must satisfy at least two requirements to ensure that prediction validity is even possible or relevant. These requirements are that (1) the intended outcomes are a function of stable individual differences to some meaningful degree and that (2) there are no organizational or work-related constraints preventing the desired outcomes. During the activity of specifying the intended uses and outcomes, inferences must be drawn from various sources of evidence confirming that these requirements are satisfied.

In addition to these two inferences, the meaningfulness of conclusions about selection validity are influenced by the comprehensiveness with which intended outcomes are specified.

A Function of Individual Differences The specified outcomes must be a function of individual differences among workers. The required inference is not so much that the eventual predictors and criteria will be measured at the level of the individual applicant and employee

but, rather, that differences among workers' outcomes are reliable and are a function of stable individual worker characteristics to some important extent.

No Organizational or Work-Related Constraints The specified outcomes must not be constrained in ways separate from the influence of individual differences that prevent worker attributes from affecting them. For example, suppose an organization proposes a selection system for an Account Rep job to increase the accuracy with which Account Reps decide whether contested charges may be removed from a customer's bill. If the current problem with inaccuracy is attributable to a lack of training about the organization's unique billing guidelines, then the intended outcome of improved accuracy is not a plausible result of any selection solution. In effect, the proposed outcome of improved accuracy does not allow a meaningful evaluation of the validity of a selection test.

These two requirements are, in effect, pre-conditions for the validity and usefulness of a selection strategy.

Comprehensiveness of Intended Outcomes The *Standards* notes that "each intended interpretation must be validated" (p. 11), but rarely, if ever in our experience, are all real benefits of a selection system ever explicitly specified as intended. The root issue is that every individual hiring decision plays some role in many positive and/or negative work behaviors and outcomes whether or not each of these real behaviors/outcomes was salient and explicitly intended in the planning of the selection system. If the employer has a broad interest in optimal hiring that selects the best employees, then, in concept, there is a very large number of "intended interpretations." For example, good employees who are contributors to the organization would be employees who work safely, show up on time every day, are helpful to others, perform their job tasks accurately and consistently at high levels of proficiency, do not steal from the organization or denigrate the organization to others, are supportive of others' success, progressively develop into positive contributors in larger and more important roles, have low health-related costs, do not leave the organization, enable others to be effective, work well in teams, take leadership roles when needed, suggest positive improvements to work processes, and so on. The point is that every hired employee produces some result on virtually every valued dimension of work behavior relevant to their work context. With rare exceptions (e.g., Project A for Army jobs; Campbell & Knapp, 2001), organizations do not specify all valued work behaviors as intended outcomes in the planning of a selection system. In fact, even when seemingly comprehensive, rigorous job analyses are conducted to empirically identify important job behaviors/results, rarely is the focus broad enough to incorporate the full range of valued behaviors, many of which extend well beyond the job itself (e.g., progression, helping, loyalty, health, and safety).

One specific example of an overlooked but valuable outcome of cognitive ability screening was reported by McCormick (2001) in a large (N = 7,764) study of the relationship between employee illness and accident rates, and cognitive ability test scores used in the employment selection process. This study found the correlation between cognitive test scores and illness and accident rates, corrected for the effects of age, to be −.07 and −.09, respectively. While these were low correlations in absolute value, the organization-wide dollar benefit of selecting applicants above the 40th percentile on cognitive ability was estimated to be approximately $96 million per year. This study describes important outcomes—health and safety behavior—that are improved by selection based on cognitive ability. While the dollar value may be surprising, it is not surprising theoretically that cognitive ability is predictive of worker health and safety behavior. Yet, in one author's (Kehoe) experience, rarely are cognitive ability tests used explicitly to influence worker health behavior. Indeed McCormick's meta-analysis was conducted across diverse selection procedures and many jobs, none of which intended cognitive ability predictors to influence health behavior. (For further information about health and safety outcomes, see Chapter 24 in this volume.)

As a very practical matter, it is common for organizations to specify only those intended outcomes that are most salient in the current local circumstances. Common examples of explicitly intended, highly salient outcomes include turnover, customer satisfaction, professionalism, sales results, speed and accuracy, project execution, share value, and revenue, among many

others. But any list of the most salient intended outcomes is virtually always an incomplete list of the real outcomes for every worker that are valued by the organization. (For further discussion about the choice of criteria, see Chapter 25 in this volume.)

Comprehensiveness in the specification of intended valued behaviors and outcomes is unattainable, as a practical matter. Further, the *Standards* does not explicitly require comprehensiveness in specifying intended interpretations. The implications are (a) validity evidence with respect to specified interpretations and outcomes is always an incomplete indicator of the relevance of a selection system to valued worker behavior and outcomes, and (b) great care should be taken at this early step in the design and implementation process to explicitly specify all the intended outcomes that the organization will expect of the selection system.

Who Specifies the Intended Outcomes? This specification of intended outcomes is critical to the design of selection systems, but there is often some ambiguity about the acceptability to various stakeholders of the specified list of explicitly intended outcomes. (This ambiguity only applies to intended outcomes; rarely is there ambiguity about the intended process(es) to be supported by a selection system, i.e., hiring, promotion, downsizing, etc.) Three sources often provide input about intended outcomes. One source is some form of standardized job/work analysis designed to identify frequent and important tasks and other work behaviors. Left to its own devices, the science-based profession of personnel selection usually begins here. However, in many cases, organization leaders (e.g., unit director, HR leader, operations manager, etc.) have strong interests in specifying the intended outcomes that are most salient and important to them. Turnover is often specified as an intended outcome in this manner. Also, more strategic outcomes may be specified by organization leaders, such as the importance of current job-specific knowledge in new hires where the organization is either increasing or decreasing its investment in new-hire job training. Indeed, in our experience it is not uncommon for organization leaders to assert that a job-specific proficiency or ability that appears important in a job analysis is, on balance, not as important as other desired outcomes and that the selection system should focus on the desired outcomes specified by the organization leader. For example, a leader of a customer service center with mostly entry-level workers may assert strongly that the most important desired outcome is reduced turnover and make the further assertion that job knowledge and learning ability are much less important. The third source of influence is the selection expert who is designing the selection system. Based on her/his expertise about personnel selection and the organization and job itself, this expert may make strong suggestions about possibly important benefits of selection outcomes that are not salient to organization leaders or identified in a job analysis.

The point of this consideration is that validity evidence will be important and useful to the organization only to the extent that the specified intended outcomes are aligned with the organization's actual interests in the selection system. Ensuring this external validity, if you will, of the specified intended outcomes is critical to the usefulness of the validation effort.

Inferences About the Quality of Predictor and Criterion Scores

No inferences about the quality of predictor and/or criterion scores follow from this first stage given that it is limited to the specification of the intended uses and outcomes. This first stage only specifies the intended uses and outcomes that are necessary in Stage 2 to identify the appropriate constructs and potential measures of predictors and criteria.

Inferences About the Prediction Rationale

The specification of the intended uses and outcomes cannot lead directly to inferences about a prediction rationale because the particular predictors and criterion measures have not yet been specified. However, the expert designer is able to use the specification of intended outcomes

to make initial judgments about the nature of criterion constructs implied by these outcome specifications. These early judgments about likely criterion constructs, in turn, enable the expert to identify potential predictor constructs from the relevant validity research foundation, but no evidence supporting a prediction rationale is gathered at this first stage.

Stage 2: Describe the Work

The process of validating selection systems depends to a great extent on the nature of the work into which applicants are selected. As a result, the information about work generated by various methods of work analysis typically provides the foundation that helps determine what many (but not all) criterion assessments should measure and, in turn, what predictor assessments should measure. For the purposes of this chapter, we treat the analysis of work very broadly to include virtually all the various processes and methods used to produce the work information required to specify criterion and predictor constructs and, in certain cases, create criterion and predictor measures. This broad treatment of work analysis includes (a) traditional job analysis methods for documenting important work behaviors, tasks, components, and required KSAOs; (b) more specialized methods for specifying work content required to develop criterion and/or predictor measures such as knowledge content, critical incidents, and situational judgments that may discriminate between good and poor judgment; and (c) methods for identifying workplace behaviors that are valued by the organization but are outside the scope of job-specific performance, such as turnover, job progression, and organization citizenship including counterproductive behavior.

As a result, the design and implementation activities that produce descriptions of work often have direct and critical influence on the specification of intended outcomes, on validity-related inferences about the quality of criterion and predictor scores, and on the prediction rationale.

Inferences About Intended Uses and Outcomes

In many cases, formal work analyses inform or specify intended outcomes beyond an initial, more general description. In this way these analyses provide information about constructs and/or content of important work outcomes that become bases for evaluating the quality of criterion and predictor measures and a prediction rationale.

Even though work analyses can further specify intended outcomes in ways that establish their importance and make them measurable, work analyses cannot replace the authority to establish the intended uses or outcomes of a selection system. Work analyses require job and testing expertise, not organizational authority. Ultimately, the establishment of intended uses and outcomes is a matter of authority, not expertise. The primary role of work analyses with respect to intended outcomes is often to further specify measurable constructs and work behavior content that capture the intention of the organization authority that initially established the intended uses and outcomes at some level of description.

Our overall perspective about the scope of work analysis is that it should be determined by the initial description of intended uses and outcomes established prior to the effort to analyze the work, and there should be no artificial methodological limits to the scope of this analysis. For example, if the organization leader prescribes that a selection system should be designed to minimize turnover, among other desired outcomes, then some analysis of the work context and conditions should be conducted, if it hasn't already, to understand the factors that influence individual decisions to leave the job or the organization. This may seem like a trite point, but the underlying principle here is that job analysis, whatever its form, should be designed to serve the purposes expressed in the intended uses and outcomes. Even in a prototypic scenario in which the selection professional has virtual free rein—within organizational constraints—to establish an optimal selection system, traditional work and worker-oriented job analyses should not be the sole determinant of intended outcomes. Because there is such a wide range of potentially important outcomes beyond the job tasks themselves (e.g., health behavior, progression success,

workplace theft, helping behavior, responsible behavior, professional/appropriate demeanor, creative/innovative behavior, safety/accident results, prosocial behavior in work teams, and early vs. late turnover), the design of the selection system should begin with some form of strategic discussion with organization leaders to identify the organization's most salient needs that are amenable to a selection solution.

Ultimately, organizational authority, well informed by selection expert information, should provide the direction needed to identify the intended uses and outcomes that will shape the selection system. The dilemma described above regarding the impracticality of incorporating all possible valued outcomes in a validation effort should be resolved by leaders with organizational authority, not by experts with job knowledge. The appropriate focus of job expertise is to identify work tasks, behaviors, and requirements that represent the intended outcomes and to help specify the content of measures of those tasks, behaviors, and requirements. The adequacy of the job expertise required for these tasks is critical to the overall claim of validity and, in particular, to the quality of predictor and criterion scores and the prediction rationale described below. (Chapter 6 in this volume provides considerably more detail about the various forms of work analysis.)

Inferences About the Quality of Predictor and Criterion Scores

Work experts produce information that can serve as construct and/or content evidence used to develop criterion and predictor measures and, in turn, to evaluate certain qualities of those measures. This expertise allows the work analysis output to be credible, which, in turn, provides a basis for claims of validity. The output of non-experts cannot provide the credibility required for any validity rationale.

An important consideration is that different types of work information may require different types of expertise. In particular, different types of expertise are required to produce credible judgments about (a) work content, (b) importance to work, and (c) assessments based on work content. Expertise in work content is required for common methods of work analysis designed to identify and describe the content of work tasks, knowledge, ability requirements, and work behaviors that constitute the full scope of work behavior relevant to the intended outcomes. This content expertise is also required to make evaluative judgments about distinctions between high and low levels of performance or work behaviors that lead to positive or negative outcomes as required, for example, in the development of job knowledge tests (JKTs), work sample tests (WSTs), and situational judgment tests (SJTs), as well as interview content that relies on critical incidents of work behavior that distinguish between successful and unsuccessful performance.

Expertise regarding the importance of work behavior, job knowledge, abilities, etc. is a different expertise than work content expertise. In many instances of work analysis, it is rightfully assumed that the same experts have both content and importance expertise. For example, where the importance of work tasks for successful performance depends on a deep understanding of the relationships of all work tasks to work performance outcomes, the expertise about importance is likely to be found in the same people who are experts about work content. However, in other work analysis tasks where the importance of tasks or knowledges depends on an understanding of organizational purposes or strategies more than an understanding or work processes, experts in work content may not be experts in work importance. For example, a call center organization may choose to place high importance on average talk time due to small profit margins, whereas call center representatives may perceive from their own experience that listening skills are more important for customer satisfaction. Where judgments of importance are required, care must be taken to ensure that experts in importance are making the judgments.

Finally, job experts are often directly involved in the development of assessment procedures including JKTs, SJTs, WSTs, and interviews. Similarly, job experts also participate in judgment processes used to develop critical test scores such as cut scores. In these cases, the judgments often require some level of expertise about behavioral assessments, which usually does not overlap with content or importance expertise. For example, using content experts to develop job knowledge items requires, among other things, that the content experts develop effective

distractors that satisfy a number of specific requirements. The most common methods by which assessment expertise is embedded in the assessment development processes used by job content experts is that standardized instructions and procedures are used and training and process oversight is provided by assessment experts.

A counterexample can be helpful. Certain cut-score-setting methods rely on job content experts to make judgments about the likely test-taking behavior of job incumbents. But job content expertise generally does provide expertise in incumbents' test-taking behavior. Perhaps the most common example is the Angoff method (1971), which requires job content experts to judge the likelihood that job incumbents who are performing at a minimally acceptable level will answer items correctly. While Angoff methods precisely describe what judgment is to be made—the likelihood of answering correctly—they generally do not choose job experts who have expertise about test-taking behavior, nor do they provide training or oversight about such test-taking behavior. In cases like this, job experts are making judgments that require an expertise they are unlikely to have. Such judgments provide no evidence supporting any validity claim for the test, nor do they provide a credible foundation for claims about the cut scores.

Inferences About the Prediction Rationale

The work description stage of design and development can provide the building blocks for a variety of prediction rationales undertaken in Stage 3 below that depend in some fashion on information about important job functions. Synthetic validity refers to a family of validation processes that rely on some judgment or index of similarity between the focal job for which a selection procedure is being used and other jobs for which empirical validity information is available from previous validity studies. Gibson and Caplinger (2007) and Hoffman et al. (2007) describe specific variations of synthetic validity evidence relating to transportability validation and job component validation, respectively, and Johnson (2007) provides an overall evaluation of synthetic validation as an acceptable technique for accumulating evidence of validity. In general, the information about the focal job used to judge its similarity to other referent jobs is generated by job experts in a structured process designed to describe the focal job in a manner that is relevant and comparable to the referent jobs. Furthermore, these same job experts or other similar job experts may also make the later judgments about the degree of similarity between jobs.

Appropriate expertise is critical for these components of prediction validation methodologies in the same manner that it is critical to conclusions about the quality of predictor and criterion measures. In all of these cases, the expertise provides the credibility of the information used to describe the links between job content and the content of criterion and predictor measures and the content of other referent jobs.

An overall observation about Stage 2 is that it provides the first place in the logical process of design and implementation where expert judgment produces information critical to subsequent inferences about the validity of criteria and predictors.

Stage 3: Choose/Develop Predictor and Criterion Assessment Processes

The purpose of this third category of design and implementation work is to specify and choose and/or develop measures of the intended outcomes and selected predictor procedures. Beyond the typical psychometric requirements for the quality of any measure, several important considerations regarding the roles of predictor and outcome measures in the validation process are described here, organized around the three categories of key validity inferences. We acknowledge that the considerations addressed here are only some of the many validity considerations relating to the quality of the predictor and criterion measures. However, the several specific considerations we describe here are among the most important and/or most contemporary.

Inferences Relating to Intended Uses and Outcomes

There are strong direct relationships between measurement quality and the intended uses and outcomes addressed in the first step of the design and implementation process. To a great extent, this is a unidirectional relationship in which earlier decisions about intended uses and outcomes and new information generated from an analysis of the target work directly inform choices about predictor and criteria constructs and assessment processes. This direct influence is an important component of the overall validity rationale for the test scores. Nevertheless, validity considerations are not the only factors in choosing among predictor options given the target outcomes. We provide the following subsection to describe the important balance between considerations of validity, the focus of this chapter, and other considerations more closely related to the effectiveness or utility of a selection system. We believe this broader perspective helps to clarify the narrower scope of validity considerations covered in the rest of this chapter.

The Roles of Validity and Usefulness in Choosing Predictors In choosing predictors for a selection system, it is obviously necessary to consider the expected validity for each potential predictor. This requires information about the outcomes (criteria) that each predictor would be intended to predict and about the accumulated research-based evidence for the validity of the predictor's scores with respect to similar outcomes. This expectation of validity is a minimum requirement for the choice of a predictor, but this consideration only serves to exclude potential predictors that do not satisfy this minimum requirement. In addition, it is critical to consider the expected usefulness of each "minimally qualified" predictor.

Evaluating this expected usefulness requires a consideration of the many complex ways in which the organization elicits valued work behavior from its employees. The selection system is just one of several parts of the whole organizational context that shapes employee work behavior. Other parts include training/development, rewards, compensation and recognition, supervisory coaching/direction, job design and supporting resources and processes, elements of organization culture that affect work behavior, recruiting sources that target particular types of applicants, the organization's reputation and attractiveness in the employment market, work governance systems such as union contracts and work rules, consequences for negative work behavior, work-life balance, inspiring and enabling leadership, and so on. Both small and large organizations can be remarkably adaptable in the ways in which they facilitate work behavior that leads to desired outcomes.

Choosing predictors requires the selection professional to consider the most useful contributions a selection system can make in this broader context. In some cases, this might also include a consideration of whether a selection solution could be a more efficient, less costly solution than the current strategy the organization uses to achieve the desired outcome. For example, a selection system might be a less costly strategy for ensuring minimum job knowledge among new employees than an early job training approach. In contrast, a cognitive ability test might add little or no value to a selection system for a computer engineering job in a highly regarded, relatively new, and successful high-tech company that attracts resumes from the top computer engineering graduates in the country. It can be instructive, if not humbling, for a selection professional to investigate the ways young, post-startup companies can be successful without adopting maximally valid, professionally designed selection practices.

The point of this comment is that maximum validity is not the selection designer's most important objective in choosing among potential predictors. The predictors that add the most value are the ones that complement (or replace) the existing organizational systems that support effective work behavior. Of course, in many cases—perhaps most cases—selection systems solve problems created by the lack of or ineffective or harmful versions of other systems supporting work behavior. In general, though, the purpose of validity evidence supporting scores on any particular selection procedure is to ensure that the specific procedure is influencing the outcome(s) as intended.

An overall point about these comments and other similar comments in this chapter is that while considerations of validity represent minimum requirements for professionally developed

selection systems, validity does not define the optimality of selection systems. Rather, organizations typically have a range of important interests that are affected by selection, and the optimality of the designed solution in any particular case is the extent to which these interests are well balanced. Validity information helps inform this balancing effort but does not define the acceptability of the various tradeoffs required to find an optimal balance.

Inferences Relating to the Quality of Predictor and Criterion Scores

This section addresses the following quality of measurement considerations: (a) validity considerations for online unproctored predictor assessment, (b) the meaning of personality scale scores across commercially available instruments, and (c) the absence of bias in criterion measures. We acknowledge that these are just three of many possible measurement quality considerations ranging from basic considerations such as reliability and item characteristics to more nuanced validity considerations such as test taker motivation and fidelity to work tasks/activities. We choose the first two considerations because they are contemporary and the professional research foundations are not settled; we choose to include criterion bias because of its sometimes subtle but critical implications for validity.

Online, Unproctored Predictor Assessment Perhaps the most significant and rapidly emerging new development in predictor assessment is online, unproctored administration. This emerging assessment methodology raises technical, psychometric, ethical, and professional practice issues that may have consequences for validity. Our overall perspective is that professional practice has evolved more rapidly than has the research foundation about the risk to validity associated with the unproctored feature of this methodology. It is widely acknowledged that unproctored administration has become a common practice (Pearlman, 2009), going so far as to make its way into mobile devices. The International Test Commission (ITC, 2006) has established practice guidelines for computer-based and Internet testing, while the more recent *Standards* (2014; Standard 10.9, p. 166) remands the practice issues for "technology-based administration" to professional judgment with no identification of issues particularly salient to unproctored online testing. SIOP's *Principles* (2003) does not specifically address unproctored or online testing but does address professional responsibility for test security and test taker identity. In this Handbook, Chapters 16, 39, and 44 address various aspects of this broad issue.

The first author's informal survey in 2013 of seven test publishers' practice of online administration of selection tests revealed large differences. Two of these publishers simply placed their paper-and-pencil assessment tools on an online administration platform with no more than one or two available forms of the tests and simply warned users that unproctored administration may corrupt the meaning of the scores. In contrast, two other publishers had developed online versions of certain tests designed specifically for unproctored administration in a manner that was largely consistent with the ITC guidelines. Key features of these tests were that (a) large banks of pre-tested items were available to enable each test taker to receive a randomized set of items with a psychometric rationale for measurement equivalence; (b) item analysis techniques and web patrols were used to proactively investigate indications of cheating; (c) users were encouraged to have a signed agreement with each test taker to adhere to the administrative instructions; and (d) proctored verification testing was recommended for short-list applicants prior to job offers. In short, while the practice of unproctored online testing is now commonplace, test publishers are widely different in the extent to which they support and encourage users to comply with ITC guidelines.

Beyond important ethical considerations (Pearlman, 2009), the impact on users for the design of selection systems and the validity of scores within those systems is that some publishers may provide no evidence supporting the psychometric test properties or predictive validity evidence of scores generated by the unproctored, online mode of administration. On the other hand, the accumulating research has generally shown that unproctored online administration leads to little, if any, variation in measurement properties (Vecchione, Alessandri, & Barbaranelli, 2012) and negligible score changes (Lievens & Burke, 2011). Similar results have been reported for measurement

invariance across mobile and non-mobile online administration with Arthur, Doverspike, Munoz, Taylor, and Carr (2014) and Illingsworth, Morelli, Scott, and Boyd (2015) showing invariance across modes for both personality and cognitive tests. However, Arthur et al. (2014) reported lower cognitive scores on mobile than non-mobile devices but similar scores on personality assessments. Overall, potentially problematic effects of lack of proctoring resulting from increased cheating do not appear to change test measurement structure or score levels for cognitive and personality assessment. However, there is some indication that mobile devices yield lower cognitive scores but not lower personality scores.

Overall, the evidence gathered to date about lack of proctoring does not show measurement or score effects that would threaten the validity of the unproctored scores. Consistent with that overall pattern of results, Kaminski and Hemingway (2009) and Delany and Pass (2005) reported no loss of validity in unproctored tests. In contrast, Weiner and Morrison (2009) reported mixed results.

Our perspective about the current state of research on the measurement and validity consequences of unproctored online assessment is that some publishers of online versions of selection tests now may have large enough databases that they can provide dependable enough measurement results to allow a local user to generalize those measurement characteristics to their local administration. However, while some publishers may have a significant amount of relevant validity data available from client users of unproctored online testing, the volume of such research published in the selection research literature is not sufficient to support broad general conclusions about the validity of unproctored scores.

Personality Scale Scores Recent trends in personality assessment research are challenging the confidence users can have in generalizations from previous research about the validity of personality scale scores to their local context. Work over the past two decades on item types that are less susceptible to faking (Stark, Chernyshenko, Drasgow, & White, 2012), ideal-point and dominance measurement models (Stark, Chernyshenko, Drasgow, & Williams, 2006), curvilinear relationships between personality scores and work behaviors (Carter et al., 2014; Le, Oh, Robbins, Remus, & Westrick, 2011), the distinctions between observer and self-report measures (Connelly & Ones, 2010; Oh, Wang, & Mount, 2011), the potential for substantive differences between alternative instruments (Davies, Connelly, Ones, & Birkland, 2015), and the stability of personality within persons over time and contexts (Green et al., 2015) have combined to limit the extent to which broad generalizations about personality validity can be made to local settings without considering specific characteristics of the setting and the personality measurement. In addition, we offer our own informal observation from reviewing dozens of commercially available personality inventories that work-specific tailored, composite scales with similar names in different instruments (e.g., team orientation, service orientation, leadership orientation) cannot be confidently assumed to measure the same facets of personality. These developments all point to the broad theme that, with regard to the generalizability of the existing research on the validity of personality scores, specificity matters far more than it does for the generalizability for cognitive test scores. The implication of this conclusion is that the selection system designer should carefully evaluate several specific considerations in establishing the local validity rationale for personality assessment. These considerations include (a) the specific workplace behaviors/outcomes to be influenced by personality assessment and the context in which these behaviors/outcomes occur, (b) the opportunity to capitalize on the potential incremental value of other-report assessments, (c) the extent to which a curvilinear (ideal point) model of assessment would be more effective, (d) the advantages of some assessments over others with regard to susceptibility to socially desirable responding, and (e) the extent to which each of several alternative assessment tools fits well with the purposes and contexts associated with the use of personality assessment.

An important implication of this increased specificity associated with the choice of and among personality assessments is the greater value (compared to general cognitive ability testing) of local criterion-oriented evidence of predictive validity.

Absence of Bias in Criterion Measures A critical concern in the process of specifying and, if needed, measuring intended outcomes is the possibility of bias in these criterion measures. One possible source of bias is the use of in-place administrative measures. Three common constraints in the validation of selection test scores are (a) the pressure to avoid costs, (b) the pressure to design and implement without delay, and (c) access only to small samples. These common constraints may collectively lead to a consideration of in-place, administrative measures of work behaviors as criteria for the purpose of validating selection test scores. Perhaps the most common of these are administrative appraisal ratings. Unfortunately, it is frequently easy to identify other factors that influence appraisal ratings beyond the target intended outcomes. These other factors may include a lack of supervisor training about the ratings, artificial distribution requirements, a lack of detailed information about actual performance, pressure to avoid low ratings that would trigger the requirement for a formal performance improvement program that supervisors might be reluctant to undertake, and other, unrelated purposes for the rating such as their use in making compensation decisions. All of these potential biasing factors are plausible threats to the meaning and fairness of appraisal ratings. For these reasons, in-place operational appraisal ratings are commonly avoided as criterion measures for validity evidence.

Inferences Relating to the Prediction Rationale

This section describes five types of validity considerations relating to the relationships between predictor scores and criterion scores.

Alignment Between Selection Procedures and Intended Outcomes The *Standards* asserts that "intended interpretations" of scores must be validated (p. 11). The implication is that it is meaningful to gather evidence of test score validity only with respect to the outcomes that are intended for that test within the design of the selection system. Consider the example of a selection system that includes a test of cognitive ability, among other things, used to hire new service reps in a call center. It is well understood in personnel selection that cognitive ability predicts task proficiency because cognitive ability enables learning of job knowledge, which is required to perform job tasks proficiently (Hunter, 1986). Here we adopt a definition of task performance from Rotundo and Sackett (2002, p. 67), "behaviors that contribute to the production of a good or the provision of a service," which is also used in a recent meta-analysis of relationship between general mental ability and nontask performance (Gonzalez-Mule, Mount, & Oh, 2014). At the same time, it is well established that cognitive ability is much less predictive and, for certain behaviors, not predictive of non-task behaviors and performance such as organization loyalty, helping behavior, citizenship behavior, and counterproductive behavior. For these reasons, the design of a service rep selection system might include a cognitive ability test to predict service rep task proficiency and some non-cognitive selection procedure(s) (e.g., personality inventory, biodata inventory, or interview assessment of team experience) to predict the desired contextual work behaviors. In this selection system, the rationales for predictive inference align the cognitive test scores with task proficiency and the non-cognitive scores with the non-task, contextual work behavior(s) of interest. The only relevant validity evidence for the cognitive and non-cognitive scores in this selection system is the evidence that is aligned with these intended interpretations (outcomes).

Two significant implications for validation follow from this "alignment" principle. First, in the case of the service rep selection system, unambiguous validity evidence is provided by correlations (or other evidence of a predictive relationship) between cognitive predictor scores and targeted task proficiency measures and between non-cognitive predictor scores and measures of the target contextual work behaviors. Correlations involving either of these predictor scores with some measure of overall performance that includes both task proficiency and contextual behavior represent ambiguous evidence of score validity within this selection system. Correlations with such multidimensional criterion measures are measures of impact or effectiveness more than they are evidence of score validity with respect to the interpretation (outcome) intended

for those scores. Correlations with multidimensional criteria that aggregate criterion measures across different types of outcomes provide ambiguous information about the intended meaning or interpretation of these scores, even though they provide very useful information about the efficacy of the selection procedures.

Complex, Multidimensional Outcomes Few, if any, work behaviors or outcomes are a function solely of the attribute(s) measured by a single selection procedure. While this condition of heterogeneous multidimensionality likely applies in virtually all cases of in-place metrics, it is certainly more severe in some cases than others. For example, the metric of "improved ROI" may be an important outcome for senior leaders, but it is certainly a highly complex, heterogeneously multidimensional outcome for which a selection test of critical thinking skills might have only a very modest influence. On the other hand, a test-based measure of training mastery may be strongly influenced by general mental ability. This point is being made about outcomes for which the heterogeneous multidimensionality is *not* a source of bias in the measure of the outcome but is an accurate representation of the causal factors influencing the outcome and the measure of the outcome. But this condition influences the evidence of validity based on measures of such outcomes. Accurate, unbiased evidence based on highly heterogeneous multidimensional outcomes will almost certainly reveal relatively low levels of validity even in the case of a highly accurate conceptual/theoretical prediction rationale. Equally accurate prediction rationales for homogeneous, more singular outcomes will likely reveal relatively high levels of validity. The implication is that the evaluation of validity evidence must take into account the complexity of the outcome as well as its measurement characteristics. Where feasible, the most theoretically meaningful validation strategy would be one in which the generality and heterogeneity of the criterion measures matches that of the predictor in question. As a practical matter, however, this is probably rarely, if ever, realized.

(Note, predictor constructs and measures have received considerable attention elsewhere in this volume especially in Chapters 11–15. In an effort to minimize overlap with those chapters, our focus in this section is on certain selected aspects of the choice and measurement of predictors that are especially relevant to evidence for the predictive validity of these scores.)

Choice Among Available Predictors The choice of predictors and associated assessment methods is critical for the design and implementation of a selection system and the accompanying validation effort. Fortunately, the profession of personnel selection has advanced to a degree that many high-quality predictor tools are commercially available with accompanying documentation of empirical psychometric and prediction evidence. This is especially the case for cognitive ability tests, personality inventories, and interview development tools. Furthermore, as this chapter is being written, online versions of these types of predictor tools have become commonplace. Overall, the implication is that now, more than before, evidence of validity for specific predictors may well include generalizations to the local setting from evidence accumulated by commercial suppliers, especially the larger consulting houses, as well as from published research.

Incremental Contributions to Overall Criterion Prediction The selection designer often has an interest in providing an overall evaluation of the validity of a set of predictor scores within a selection system. A frequent strategy used in the selection profession to describe the validity of a set of predictors is to regress a measure of overall performance on those multiple predictors and report the increment in the multiple R^2 attributable to each predictor. Schmidt and Hunter (1998) provide a well-known, high-level example of this type of analysis. Although this analysis can have useful heuristic value, it has two significant limitations as a form of validity evidence for the specific predictors. First, this approach relies on the construction of the overall, complex criterion measure that is a weighted composite of all the outcomes that were explicitly intended for each of the predictors. Second, even if the overall measure captures all intended outcomes, this regression analysis produces coefficients (multiple Rs) that are influenced by the relative weighting and interrelatedness of the multiple outcomes in the construction of the overall criterion measure. The net consequence of these two limitations is that the meaning of multiple Rs has more in common with utility analysis than with validity analysis, where utility

analysis focuses on a magnitude of relationships and validity analysis focuses on the meaning of relationships. Multiple R and the increase in R (or R^2) do not provide unambiguous evidence of the extent to which scores on each of the predictors is predictive of the outcomes it was designed to predict (meaning). This is not a criticism of this type of regression analysis. Rather, it is a cautionary note that this type of evidence has a different meaning than evidence of the relationship between a predictor and its intended outcomes. For example, the usual result that personality predictors contribute less variance than cognitive predictors to overall performance measures does not necessarily imply that personality scores are less valid predictors of their intended outcomes than cognitive scores are predictive of their intended outcomes. (Of course, we know from separate validity evidence that personality scores generally do correlate less with the work outcomes they are conceptually expected to predict than do cognitive predictors.)

One implication of this comment is that the question of incrementalism with respect to selection procedures within a selection system is, at root, a question of value or utility and is not an unambiguous indicator of validity. One can easily imagine validity evidence being used to determine whether one predictor is more or less valid with respect to its intended outcomes than another predictor is of its own intended outcomes. But as soon as the question is about the *incremental* value of one predictor with respect to another, the question fundamentally hinges on, among other things, the relative value to the organization of the two sets of intended outcomes, which is independent of the question of validity.

Generalizing Validity Conclusions from Previous Research Criteria to Criteria in the Local Setting Given common constraints on (a) the cost and time available to design and implement selection systems, (b) limited local sample sizes for local empirical studies, and (c) the challenges of accurately measuring the intended outcomes, an increasingly common and effective validation effort relies on generalizing conclusions from previous validity research to the local setting. Indeed, it seems likely that at least some part of the validation rationale for every local selection system relies on some generalization from previous research conclusions to the local context. Beyond the ordinary psychometric requirements for criterion measures, we make three points here about conclusions about local criterion measures based on previous research conclusions. First, the constructs captured by local criterion measures will be specific to the local context in virtually all cases. For example, even though "turnover" is a generic label for a common type of criterion measure, the meaning of a local measure of turnover—as a criterion to be predicted—is likely to be highly contextual given the particular factors causing local turnover. Similarly, a properly instructed supervisory rating of local, overall job performance will capture the facets of job performance important in the local job (Campbell, 2015). Second, the inference that conclusions about criteria from previous research apply to local criteria will be based on the conceptual similarity between the constructs underlying previous research criteria and local criteria and will not be based on any type of sampling rationale. Third, it is highly likely that the inference of conceptual similarity between previous and local criteria will be based on expert judgment rather than on some quantifiable comparison algorithm or on the use of identical measurement procedures.

Summary conclusions about the criteria represented in meta-analytic research studies that include several local studies will often be at a different level of description than the local criteria. Conclusions from such cross-study research efforts will typically classify or categorize the studied criteria in an attempt to reach a more general conclusion. As a result, the inference that previous validity conclusions for categories of criteria can be generalized to a local criterion requires the expert judge to evaluate whether the locally specific criterion constructs and measures are similar enough to the constructs and measures captured by research-based categories of criteria.

The expertise involved in this judgment should include knowledge about the general principles of inference and measurement as well as knowledge about the substantive meaning of the criterion constructs and measures in the previous research and in the local setting. We make this point here to underscore the importance of expert judgment in reaching a conclusion about test score validity in a local selection system. The role of expert judgment in generating validity evidence is well-established in professional guidance and in practice. The *Standards* frequently cite and endorse the role of expert judgment as a source of validity evidence. (See, for example,

Standards 11.3 and 11.5 and their accompanying comments). Gibson and Caplinger (2007) and Hoffman, Rashkovsky, and D'Egidio (2007) describe the roles of expert judgment in a variety of structured methods such as job component validation for drawing inferences about local validity from previous validity evidence. We single out the role of expert judgment here because it is likely to take on even greater importance in establishing the local validity evidence where a local criterion study is not feasible.

Stage 4: Prescribe Score Usage

A critical consideration in the design of a selection system is the manner in which test scores will be used in the process of selecting among the applicants. The irony is that, as critical as this design component is for the effectiveness of the selection system, with a few exceptions it has relatively little consequence for the type of validity evidence to be gathered for the predictors. In this section we first consider three types of score usage that may have implications for the nature of the appropriate validity evidence, and then we briefly describe a systematic approach to the design of selection systems that is consistent with an overall theme in this chapter that validity is a critical building block but does not define the optimality of a selection system.

Selection designers have many options available to them regarding the manner in which test scores may be used: (a) scores may be used in a compensatory or non-compensatory fashion; (b) scores may be used in a wide variety of ways to establish selection standards in the form of cut scores and/or score ranges associated with specific decisions; (c) scores may be used to screen applicants in a particular sequence; (d) scores may be used to inform individuals who make the selection decisions with certain guidance accompanying the score information; and (e) scores may be weighted to control their relative influence on selection decisions. Of all the ways scores may be used, only three of these ways have implications for needed validity evidence. The nature of the required validity evidence will be influenced by the choices about (a) weighting predictor scores in some form of compensatory scoring, (b) the manner in which scores are used to inform selection decision makers, and (c) the sequence in which scores are used to affect selection decisions. These are described below in the section Inferences Relating to Quality of Predictor and Criterion Scores and in the section on Inferences Relating to the Prediction Rationale. However, before addressing these three key issues, we first address considerations relating to intended uses and outcomes.

Inferences Relating to Intended Uses and Outcomes

None of the myriad ways of using scores is likely to have implications for validity evidence relating to intended uses (e.g., hiring vs. training admissions) or intended outcomes. This is because the manner of score use has no necessary consequences for the intended uses or outcomes. Most decisions about score use are driven by considerations of operational efficiency or feasibility and do not alter the validity rationale required to support the intended uses and outcomes from the scores. A common example is the choice between use of some form of cut score–based strategy rather than some alternative such as top-down selection. In this case, the common professional practice is to comparatively evaluate these alternative uses by analyzing their implications for cost, efficiency, diversity, risk of adverse impact, and, possibly, other consequences. But this comparative evaluation ordinarily does not assume or estimate different validities for the same selection procedure used in these different ways.

(Note, we acknowledge here, in anticipation of points made below, that a persuasive argument can be made that the validity of dichotomized test scores, as would be used in effect by certain cut score strategies, should be estimated separately from the validity of scores used in their original, more continuous scale form, as might be the case with top-down selection. This is an argument that scores should be validated *as used* to make selection decisions. We will revisit this argument below in the cases of composite scores and scores used to inform hiring manager decisions.)

Inferences Relating to Quality of Predictor and Criterion Scores

Scores Used to Inform Selection Decision Makers' Judgments A common use of test scores is to present them to selection decision makers in some organized fashion that helps the decision maker integrate the meaning of the score information with other applicant information to form a judgment about the applicant's overall quality. Setting aside the well-established point that human judgment tends to suboptimize the aggregation of valid score information, this use of scores would require new validity evidence supporting the manager's judgment if the constructs captured in that judgment are different in some fashion from the original scores and, therefore, required a different prediction rationale. We recognize that this is a debatable claim, but we argue that in those settings in which the selection designer chooses to use test scores in this fashion, this design decision often rests on a belief (held by either the designer or the decision maker) that the decision maker has additional relevant information that improves on the test scores and makes better selection decisions. In this case, the decision maker's judgment about the applicant represents a new measure of different predictor constructs and, therefore, requires separate evidence of validity beyond the evidence for each contributing test score. We also recognize that this conclusion can be problematic as a practical matter because, often, the only output capturing the manager's judgment is the set of selection decisions. This precludes any validity evidence that depends on differentiation among the selected applicants. In those cases where the decision maker's judgment cannot be captured in an overall rating or ratings of specific applicant attributes, it will probably not be feasible to gather empirical evidence of predictive validity.

Inferences Relating to the Prediction Rationale

Here we address the implications for validity of two types of score usage—composite scores used in compensatory approaches and scores used in sequence in multiple hurdles approaches. In both cases, we describe ways in which these two methods of score use change some feature of the prediction rationale and, as a result, change the type of validity evidence required to support scores used in that manner.

Score Weights to Produce Composite Scores Compensatory scoring requires that composite scores be arithmetically derived from individual test scores. Composite scores will require additional validity information beyond that required of the individual components where these composite scores represent a new measure relating to intended outcomes differently than the component test scores used to form the composite. This will occur when the component weights used to form the composite are intended to be a measure of the relative importance of the predictor constructs for successful job performance. In this case, the composite includes additional information—the importance weights—beyond the information in the separate components, so the new information is justified based on its job relevance. As a result, validity evidence for this type of composite is supported by some rationale for the job relevance of the weights.

Similarly, setting aside the issue of weighting, it is conceivable that a composite score has different meaning than the simple sum of component scores if the attributes represented by the component scores are interactive in such a way that particular profiles of component scores have predictive meaning unique to the particular profiles. For example, considering personality attributes, if applicants who are 1.5 SDs above average on narcissism are predicted to be poor performers regardless of other attributes but applicants who are 1.5 SDs above average on assertiveness are predicted to be poor performers to the extent that they lack other important attributes, then any linear combination of component scores would likely be less predictive of performance than the component scores used individually. (Of course, in this case, a composite score would be a poor choice for this very reason, so the question of whether it would warrant separate validity evidence would be moot.)

On the other hand, the predictive rationale underlying a composite of test scores requires no new theoretical or conceptual consideration or new validity evidence where the weights are not based on any job-related consideration and where the component attributes do not have an interactive relationship to the target outcome.

It is worth noting that this same rationale may be applied to other ways of using scores such as cut scores. Cut scores are rarely, if ever, based solely on a prediction rationale linking particular scores to a job relevance interpretation such as a minimally acceptable level of performance. Rather, they are often based on a set of considerations relating to cost, manageability, optimized effectiveness, group differences, and the like. For this reason, particular cut scores rarely, if ever, rely on a claim of validity other than the fundamental claim that cut scores are based on valid scores.

Sequential Use of Scores The implication for validity evidence of the decision to use scores in a sequence is a technical point relating to range restriction. This point applies to quantitative evidence of validity in the form of correlations between test scores and outcomes among those who are selected into that stage. At each stage, a particular test score or composite of test scores is used to make selection decisions about which applicants move to the next stage. Consider an example of a two-stage sequence in which, at Stage 1, 50% of the applicants are screened out based on a cognitive ability test score. At Stage 2, the surviving applicants are given a personality assessment of Openness and 50% of them are screened out based on their Openness score. The surviving applicants are then given job offers. Subsequently, the new hires' relevant work outcomes are measured and correlations are computed between the cognitive scores and the outcomes intended from cognitive test and between Openness scores and the outcomes intended from the Openness inventory. Because these two correlations were computed only among the new hires, they are artificially range-restricted estimates because the ranges of cognitive scores and Openness scores among the new hires are both less than the ranges of cognitive scores at Stage 1 and Openness scores at Stage 2. Both estimates should be corrected for range restriction but in different ways. The correction of the cognitive validity coefficient should be with respect to the range of cognitive scores at Stage 1, whereas the correction for the Openness coefficient should be with respect to the range of Openness scores at Stage 2.

(Note, in this example the Openness scores in Stage 2 may have been indirectly restricted by the selection on cognitive scores in Stage 1. This is because Openness and cognitive ability typically are positively correlated, but this indirect restriction between stages is immaterial to the method used to correct the restricted validity coefficient for Openness at Stage 2. For Openness, the restricted validity coefficient computed among new hires is corrected for the range of Openness scores observed at Stage 2 regardless of the role of indirect restriction due to screening on cognitive scores. However, this Openness validity coefficient corrected for the range restriction among new hires should be interpreted as an estimate of the predictive validity of Openness scores where preselection on cognitive ability has taken place. This estimate of the validity of Openness scores is not generalizable, without further correction, to a different local setting in which Openness scores are used for selection from an applicant pool that has not been prescreened on cognitive ability.)

Other than these three specific uses of test scores, we believe that no other manner of score use affects the type of validity evidence appropriate to the test scores.

Stage 5: Prescribe Governing Policies and Rules

Virtually all selection systems are shaped and governed by a set of policies and rules. These typically address many facets of the selection system ranging from applicants' access to the selection process, management of applicant data, and the permissibility of waivers and exemptions to testing processes such as applicants' option to retake a test, accommodations in the testing process, and permitted modes of administration. Detailed descriptions of such policies and rules are presented elsewhere (e.g., Kehoe, Brown, & Hoffman, 2012; Roe, 2005; Tippins, 2002, 2012)

and in Chapter 9 of this volume. The focus in this chapter is on those policies and rules that can have implications for the meaning of and evidence of validity for test scores.

We briefly consider the validity implications for policies relating to retesting, mode of administration, equivalencies, accommodations for disabilities, test preparation, and exemptions and waivers. Each of these practices, except for exemptions and waivers, can affect test scores. As a result, it is important to consider whether they trigger the need to gather different types of validity evidence. In our analyses of the validity implications of these policy-driven practices, we rely on the distinction between standard and non-standard administrations of tests in the selection process and acknowledge that feasibility and practical impact are major considerations.

Inferences Relating to Intended Uses and Outcomes

The two policies addressed here affect intended outcomes but do not require any additional type of validity evidence. These two policies are about (a) equivalencies and (b) exemptions and waivers. We describe these here to document examples of selection system policies that do not warrant unique validity evidence.

Equivalencies Some selection systems establish "equivalency" rules or standards by which some other attribute of an applicant may be treated as equivalent to a test result, and the applicant is given the same status that would have been earned from the test result. For example, a personality score for Service Orientation (the referent test) is used as a requirement for several different types of customer service jobs. External applicants and internal employees may apply for these jobs. The organization has a policy that internal applicants who have a supervisor's rating of, say, 3 or higher on a standard organizational competency of "Works Well with Others" will be assigned a "passing" score result on the Service Orientation measure. In this example, applicants have been granted a score status on a selection test they have not taken because some other result—in this case, a performance rating—is interpreted by the organization as having a comparable predictive value for an intended outcome.

The question we raise here is whether the set of test scores used to validate the referent test in this example should include "awarded" test scores assigned to certain applicants via the equivalency policy. In our view, no, these awarded scores are an administrative vehicle for giving applicants a qualification status based on other considerations relating to perceived acceptability, efficiency, and fairness as well as a plausible professional judgment of comparability of meaning. The awarded statuses are not intended to be interpreted as having the same meaning as the referent test scores but are intended to be interpreted as having similar enough meaning to warrant giving the applicant the awarded qualification status. In this case, the estimate of the validity of the referent test scores would not be more accurate by including the awarded score results in the local validity study.

Exemptions and Waivers Equivalency policies describe multiple ways in which a test score result may be awarded, including completing the test. In contrast, exemption and waiver policies describe certain circumstances in which an assigned authority may decide that an applicant is not required to satisfy one or more standard job qualification requirements. For example, consider a selection process for account executives that requires satisfactory performance on a sales assessment work sample test. An organization may choose to exempt applicants from this sales assessment requirement who have been deliberately recruited from a competitor's account executive role. In our experience, exemption and waiver policies are quite common, even if they are quiet or implicit or have a much less relevant rationale than the account executive sample in which previous experience was a justification for the exemption. While the liberal use of exemptions and waivers can harm (or help) the effectiveness of the whole selection system, they do not have any implications for the appropriate validity evidence for the exempted selection procedure. That is, no validity claim about the selection procedure is strengthened by gathering evidence that the exempted applicants are as likely as high-scoring applicants to produce the intended outcome. Indeed, once the exemption authority has been established, it isn't necessarily the case

that exemption decisions must be based on expected performance. Other personal or organizational considerations may be the bases for exempting certain applicants from a standard selection requirement. In short, there is little to be gained by having the validation rationale for a selection test take into account those instances in which the test requirement is waived.

Inferences Relating to Quality of Predictor and Criterion Scores

Three common and important policies about retesting, test preparation, and mode of administration are known to have direct consequences for test scores but do not depend on any change in the predictive rationale for the target tests. The sections below explore the implications of these policies for unique validity evidence that may be warranted.

Retesting It is a common practice in selection systems to allow applicants to retake selection tests, guided by organization policies. For example, a typical requirement is that applicants may retake a test only after waiting for a prescribed period of time, which may vary by type of test. Considerable research has investigated the effects of retesting on cognitive and non-cognitive test scores. For cognitively loaded tests, evidence shows that second occasion scores are approximately .25–.50 SDs higher than first occasion scores (e.g., Hausknecht, Halpert, Di Paulo, & Moriarty Gerrard, 2007; Lievens, Buyse, & Sackett, 2005), but findings are mixed regarding criterion validity differences and measurement equivalence between first and second scores (e.g., Lievens, Buyse, & Sackett, 2005; Lievens, Reeve, & Heggestad, 2007; Van Iddekinge, Morgeson, Schleicher, & Campion, 2011; Villado, Randle, & Zimmer, 2016). The very large and persuasive body of empirical research (e.g., Schmidt & Hunter, 1998) showing (a) substantial predictive validity for cognitive tests with respect to job proficiency criteria and (b) low variability in predictive validity across a wide range of jobs and settings is generally regarded as persuasive evidence that professionally developed cognitive tests will have substantial predictive validity with respect to proficiency criteria in local settings. As a practical matter, this inference of predictive validity also is relied upon to assure selection designers that retest effects are not important sources of invalidity, even if some studies have shown changes in validity and measurement structure with second test scores.

For (non-cognitive) personality tests, the retesting issue is quite different theoretically, empirically, and in practice. The dominant theoretical consideration is about the susceptibility of self-reported personality scores to be intentionally influenced by impression management, or faking, as it is frequently called. Significant research has been conducted to understand and estimate the effect size of faking as a source of construct invalidity (e.g., Hogan, Barrett, & Hogan, 2007; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Ones, Viswesvaran, & Reese, 1996). The effects of faking are a dominant consideration in research on personality retesting for two reasons. First, a common research paradigm that investigates retest scores is one in which the study participants who have retaken a personality assessment are doing so because they were not hired following their first attempt. Second, unlike cognitive retesting, there is little theoretical rationale that could attribute score changes across short retest intervals to development or growth in the target personality attributes. Rather, the more compelling theoretical rationale to possibly explain personality score changes is that the test takers are motivated by their initial failure to adopt a different model of the personality attributes presumed to be desired by the employer. As a result, the most salient factor in attempting to explain personality test-retest score differences is faking. Unlike cognitive tests, it appears to be generally accepted that changes in personality scores from first to second scores are faking and random error, both of which introduce invalid variance.

Two threads of empirical evidence are important as sources of evidence for the validity of retest personality assessments. One thread (e.g., Hogan, Barrett, & Hogan, 2007; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Ones, Viswesvaran, & Reiss, 1996) focuses on differences in predictive validity correlations between first occasion scores and second occasion scores. This thread of research has found that, overall, the predictive validities for first scores and second scores are similar. The second thread focuses on the differences between the distributions of the

first and second occasion scores. In particular, this line of research has investigated the extent to which the selection decisions are different when an applicant pool includes only first occasion scores as compared to applicant pools that include both first occasion scores and retest scores (e.g., Walmsley & Sackett, 2013). In general, this research has shown that the inclusion of higher retest scores, which are not uncommon (Hausknecht, 2010), can significantly change who is hired. This result, if generalizable to local settings, can be taken to mean that the policy allowing retesting for personality assessments may reduce the effectiveness or efficiency of a selection system, even if original and second scores are approximately equally valid predictors, by increasing the percentage of new hires who benefited from a score increase that is construct invalid to some extent.

The practical consequences of the theoretical and empirical state of personality assessment are complex. Design decisions about personality assessments vary considerably, although it is clear from the large number of commercially available tools that personality is a common component of current selection systems. The first author's personal experience indicates that (a) retesting is probably commonplace, (b) corrections for faking and/or retest effects are probably not commonplace, and (c) local job conditions and important outcomes are probably important factors, more so than with cognitive tests, in the decisions about which scales and associated instruments are used. It would be difficult to gather unambiguous validity evidence in support of decisions (a) and (b), so pragmatic considerations relating to applicant perceptions and satisfaction and to process efficiency and cost are likely the most important considerations. Design decisions relating to (c), however, may be informed by the considerable published evidence (from publishers and from professional research efforts) about the specificity of personality scales' predictive validity, other than the evidence for Conscientiousness, which generalizes across a wide range of jobs and work behaviors.

In general, for both cognitive and non-cognitive tests, the degree of uncertainty about test-retest score equivalence and predictive validity is regarded as an acceptable risk in the design of selection systems for two reasons. First, for specific issues like test-retest practices, current research conclusions are not clear enough, except for expected differences in cognitive scores, to generalize research-based conclusions to a particular setting. Second, it is probably not feasible in most cases to conduct local test-retest predictive validity studies.

Test Preparation The *Standards*, Standard 8.0, asserts that test takers have the right to, among other things, "adequate information to help them properly prepare for a test." This standard is based on the guiding principle that test takers have a right to be informed and on the underlying belief that proper preparation enables test takers' test performance to more accurately reflect their standing on the tested attribute. The net effect of this professional standard is that selection system designers are unlikely to seek out validity evidence relating to specific test preparation policies, with the exception that such policies should avoid inappropriate preparation practices.

While we are not aware of surveys describing the current state of practice with regard to test preparation, it is likely to be common, especially for high-volume tests for which there is a "market" for test preparation courses and materials. Test preparation ranges from basic information about the test purpose and item format(s) to access to practice versions of similar tests and to more detailed instructions about processes for finding answers to items and opportunities to practice with feedback (Sackett, Burris, & Ryan, 1989). We also anticipate that the increased use of online test administration, especially mobile applications, will lead to a significant *reduction* in test preparation resources that are available on the same media and platforms as the test. (Note, changes in the frequency with which mobile devices are used may occur rapidly. However, recent large studies of unproctored online test usage reported that only 1%–2% of online test takers used mobile devices (Arthur, Doverspike, Munoz, Taylor, & Carr, 2014; Illingworth, Morelli, Scott, & Boyd, 2015).)

Test preparation has much in common with retesting both conceptually and empirically. Indeed, studies of test preparation and practice effects often treat retesting as a form of practice, especially in the case of cognitive tests. Studies of the effects of test preparation on cognitive test scores show overall very similar effects to retesting (Hausknecht, Halpert, Di Paulo, & Moriarty Gerrard, 2007; Kulik, Bangert-Drowns, & Kulik, 1984; Lievens, Buyse, Sackett, & Connelly, 2012; Lievens, Reeve, & Heggestad, 2007). Indeed, retesting in the form of multiple practice

tests is considered a form of test preparation. Nevertheless, studies also show that type and amount of preparation/coaching can affect scores differently (Kulik, Bangert-Drowns, & Kulik, 1984; Powers, 1986).

For non-cognitive assessments such as personality inventories, test preparation may consist only of examples of the item types and formats to reduce the novelty of these inventories and clarify the meaning of the instructions.

Test preparation for interviews supports an entire cottage industry and is often out of scope for the employer. Rather, various support organizations such as schools, private sector companies, unions, and search firms are far more likely than employers to offer interview preparation programs for job seekers. Similarly, preparation for physical ability testing is often supported by applicant support groups such as schools and unions.

In practice, test preparation policies are likely to consider a much wider range of possible practices than are considered with regard to retesting. Perhaps the two most common considerations regarding test preparation are cost and appropriateness. Generally, cost considerations are treated as having little, if any, relevance to issues of validity, even though it is quite likely that more costly test preparation programs such as extensive study materials and access to practice tests may lead to larger score increases than less costly programs such as pre-assessment instructions about test taking and exposure to sample items (Powers, 1986).

Appropriateness considerations, on the other hand, are often regarded as being directly related to score validity, especially for skill, knowledge, and ability tests. In these cases, it is generally regarded as inappropriate to provide parallel tests as practice forms and to provide item-specific instructions that teach the knowledge, skill, or ability being tested. Such test preparation strategies that are so "close" to the operational test and items are inappropriate because they presumably lead to artificial, invalid score increases that reflect newly learned test/item-specific information without enhancing the target construct. This is personnel selection's version of education's "teaching to the test" problem.

Mode of Administration Many of the considerations relating to the consequences of online test administration for test validity were reviewed above. The one point to be made here in this discussion of policy implications for validity is that traditionally the science-oriented practice of selection testing placed great emphasis on a consistent, standardized mode of administration. Indeed, this emphasis is sustained in the current *Standards*. For example, Standard 6.1 regarding Test Administration reads, in part, "Test administrators should follow carefully the standardized procedures for administration and scoring specified by the developer." It is difficult to imagine how unproctored online test administration can even remotely comply with this Standard. First, there is no test administrator with an implied enforcement role to ensure adherence and consistency. Second, the understood meaning of standardization that all test takers complete the test under the same, beneficial administration conditions is conspicuously violated. In this now frequent context, a new burden falls to test publishers and users to demonstrate that a conspicuous lack of standardization does no harm to test score validity, but it is difficult to conduct research about unstandardized practices with sufficient controls to justify clearly prescribed, generalizable conclusions. The profession is left with the approach recently seen in which online score results are investigated in huge data sets. Arthur, Doverspike, Munoz, Taylor, and Carr (2014) reported analyses of more than 3.7 million applicants who completed online tests. Illingworth, Morelli, Scott, and Boyd (2015) reported analysis of more than 935,000 applicants who completed online tests. This approach invites the user to conclude that whatever the online administration circumstances are in the local setting, they are captured by the mega databases that report negligible score change and measurement invariance across modes of online administration. However, the actual profile of variations in administration represented by the mega "sample" cannot be specified because the information isn't available. For example, what percentage of test takers in the mega samples attempted to cheat? What percentage were not who they said they were? What percentage were online savvy? What percentage had completed online tests before? What percentage weren't motivated to perform well? And so on. The sheer size of these mega samples does not ensure that any particular local applicant pool will produce similar results because it is impossible to know what characteristics of the local applicant pool

and their use of online administration options matters with regard to score meaning and predictive validity but may have been completely obscured by such large samples.

A consequence of this current status is that a greater burden is placed on the selection system designer to evaluate plausible local threats to the meaning and predictive validity of scores from online administration.

Inferences Relating to the Prediction Rationale

Policies relating to applicants with disabilities are addressed in this section because the most salient feature of these policies is the extent to which the federal regulations that implement the Americans with Disabilities Act (1990) (ADA) requires employers to ignore the significant loss of prediction rationale likely caused by such accommodations.

It is likely that the large majority of medium to large organizations have established some policy relating to testing applicants with disabilities. These policies may cover a variety of aspects of the selection process, including the manner in which disabilities are disclosed, the organization's responsibility to consider reasonable accommodations in the selection process, and the bases for deciding what actions to take in response to applicants' requests. Further, the ADA obligates employees to make individualized decisions about accommodations. The result of all these considerations is that organizations may offer some form of individualized accommodation to one or more aspects of the selection process. This often leads to a set of circumstances in which tests are administered using accommodated processes, scores are recorded and relied on for selection decision making despite the likelihood that little, if any, information—including validity evidence—is available to support the rationale that the scores predict the desired outcomes. The considerations for validity are unique with virtually no parallel in employment selection.

Accommodations for Disabilities Under ADA and the ADA Amendments Act of 2008 (ADAAA), employers have an obligation to consider and provide reasonable accommodations to disabled applicants in the work setting as well as in the selection process. Campbell and Reilly (2000) and Guttman (2012) provide detailed descriptions of employers' legal obligations and of common and accepted practices for accommodating disabilities in the selection process. In addition, Campbell and Reilly summarize the scant empirical evidence about the effects of disability accommodations on test scores and predictive validity. We do not reiterate those summaries here. Rather, we focus on the central validity issue posed by the legal obligations ADA imposes for disability accommodations. That validity question is whether accommodated test scores are predictive of the disabled person's performance of "essential job functions," while eliminating an artificial bias in unaccommodated scores that would lead to under prediction of such performance. For example, does an accommodation for visual impairment that uses large print materials both eliminate an under prediction bias and yield test scores that are predictive of performance of essential functions?

What validity evidence can be available to employers to evaluate this validity question? Certainly, it is difficult to justify generalizing previous validity conclusions from standardized administration processes to the local scores from non-standardized accommodated administrations. And, local empirical validation studies are almost always infeasible simply because of the low numbers of applicants who disclose the same particular disability and receive the same accommodation. As a result, employers will rarely have the opportunity to rely on meaningful empirical evidence either from previous studies or from local studies. In virtually all cases, employers must rely on the expert judgment of the test developer (or some expert surrogate for the developer) to evaluate the theoretical and empirical bases for concluding that accommodated scores both eliminate bias and are predictive of essential function performance. It is important to note here that this reliance on expert judgment is not unique to disability accommodations but, in fact, is quite common where persuasive local empirical studies cannot be conducted. Synthetic validity strategies and generalizations from meta-analytic studies rely on the same expert judgment about the persuasiveness of the theoretical and/or empirical rationale that local scores will be unbiased and predict local performance.

Stage 6: Manage and Maintain the Selection System

This section addresses four elements of managing and maintaining the selection system: (a) training selection administration staff for operational knowledge and skills, (b) auditing for compliance with policies and processes and for indicators of threats to effectiveness and validity, (c) adapting the operation of the system to changing needs or circumstances, and (d) maintaining current professional expertise. Kehoe, Mol, and Anderson (Chapter 9 in this volume) provide a more broadly focused summary of managing for sustainability over time.

We note that the four elements of these maintenance practices described here can have implications for all three categories of validity inferences. Nevertheless, it is useful to align these four elements with the categories of inference they are most likely to influence.

Inferences Relating to Intended Uses and Outcomes

Adapting Our primary point with regard to readiness to adapt is that even though selection systems are rooted in stable individual differences that reliably shape important work behaviors, a variety of organizational and professional changes may create the need to change a selection system. A compelling example is the rapid impact of the availability of online assessment tools to provide less expensive and faster selection processes.

While our comments regarding adapting are much less prescriptive than for training and auditing, we suggest the following management strategies for recognizing, evaluating, and adapting to organizational and professional changes that point to improvements in an existing selection system:

1. Periodic reviews with unit-level HR leaders can be a very effective strategy for establishing access to information about organization changes that might have implications for selection.
2. To the extent possible, selection leaders should capitalize on the data described in the Auditing section below to become the owners and producers of periodic reports to organization leaders that convey the linkage between work behavior outcomes and selection processes. Treating selection scores as metrics of an organizational process, and linking them to outcome measures, positions the selection scores and the selection leader as credible and valuable sources of information about important outcomes.
3. Understanding validity as a means to an end, which is a major theme of this chapter, is a professional perspective that is likely to create more openness to view validation processes as a large toolkit of methods and procedures, some of which are more suited to current local circumstances than others.

Inferences Relating to Quality of Predictor and Criterion Scores

Training The training of selection system staff and role players is important to maintain validity because the quality of scores and the appropriateness of selection decisions depends on the successful performance of several peripheral functions, including applicant recruiting, interviewing, administration and scoring of tests, the processes of properly relying on selection scores to help make the intended selection decisions, the processes of creating and maintaining effective applicant management systems, and the management of accurate databases of score results and other applicant information.

We describe three recommendations to optimize the benefits of training for the maintenance of valid and effective selection systems:

1. Provide training for all functions that are critical to a well-managed selection processes.
2. Develop training processes that require trainees to demonstrate minimally effective skills in order to be certified in the target function. Certify successful trainees.
3. Require that all critical functions are performed only by people who are training certified.

We recognize that these three requirements may collectively be an onerous requirement, and the organization may need to adopt an approach that allows it to gradually achieve this objective, but it is important to acknowledge their importance for sustaining selection validity and effectiveness.

Auditing An auditing function is central to the management and maintenance of selection validity and effectiveness. Perhaps the greatest operational threat to test validity over time is the gradual loss of discipline and adherence to the process requirements for effective and valid selection procedures. Coupled with staff training, an effective auditing function can help maintain disciplined adherence to appropriate processes in two ways. First, auditing signals to the staff and stakeholders that disciplined adherence is critical. Second, auditing provides information about key indicators of process adherence and selection outcomes. Overall, effective auditing should provide at least three types of information about selection systems: (a) periodic evaluation of score properties, (b) continual confirmation of process adherence, and (c) periodic data about the achievement of the intended outcomes.

Inferences Relating to the Prediction Rationale

The effective management of selection systems influences the soundness of prediction rationales primarily through the effort to sustain a high level of expertise in the selection professionals who support the organization. Expertise has two primary roles in maintaining valid and effective selection systems. First, professional expertise is a frequent source of evidence supporting claims of validity by providing expert judgements about job tasks and requirements. Second, professional expertise about the research foundations and professionally developed tools and resources may recognize new solutions to organization priorities and needs.

SUMMARY OF PART 2

Part 2 of this chapter proposes a six-stage process for the design and implementation of selection systems and describes significant considerations in each stage that can have implications for validity inferences about (a) the intended uses and outcomes, (b) the quality of predictor and criterion measures, and (c) the prediction rationale. The six stages are described in a logical order from (1) specifying the intended uses and outcomes to ensure the outcomes are amenable to a selection system based on stable individual differences in work behavior, (2) describing the work in a manner that identifies work content and its importance to inform decisions about locally relevant predictors and criteria and supports inferences (decisions) that conclusions from previous research apply locally, (3) choosing and/or developing predictor and criterion measures based on clear understanding of the likely causal linkage between test scores and work behaviors and outcomes, (4) prescribing the manner in which predictor scores will be used that capitalizes on the causal linkage while accommodating local constraints, (5) prescribing the policies and rules that govern the selection system to ensure its validity and usefulness across all conditions, and (6) managing and maintaining the selection system to control or adapt to the dynamic factors that can change validity and usefulness. At each stage of work, information is generated and inferences (decisions) are made that strengthen or weaken the claim of predictive validity. The overall claim of selection system validity can be represented as a conclusion based on the aggregation of many diverse sources of empirical and expert evidence accumulate across the design and implementation stages of work. We believe this way of describing validation contributes to our professional understanding of the meaning of validity, the distinction between validity and effectiveness, and validity's role in the selection professional's effort to provide useful methods for achieving individual and organizational success.

CONCLUSIONS

This chapter explored our professional understanding of selection validity and examined how the design and implementation of selection systems generates information needed to support, ultimately, the claim of predictive validity. Several key conclusions emerged:

- Evidence supports the validity of test scores when it supports the claim that intended outcomes follow from the use-specific interpretation of test scores. Other evidence about the effectiveness and value of selection systems may be critically important and, possibly, more important, but only evidence relating to the meaning of the test scores supports claims of validity.
- Decisions made throughout the design and implementation process are often inferences made by the selection expert that empirical evidence gathered in previous validity research efforts generalizes to the local setting.
- Factors affecting score validity are dynamic and must be managed with regular auditing processes.
- Expert judgment is a critical source of evidence for the local validity of test scores and, in some cases, may be the primary source.
- Not all decisions about selection tests depend on or produce validity evidence. For example, decisions about the manner in which test scores are used (e.g., cut scores, advisory input, banding) and decisions about governing policies such as exemptions and waivers often do not require validity evidence but, instead, must be supported by evidence that the expected outcomes will be consistent with organization requirements such as speed, cost, efficiency, user satisfaction, and degree of improvement in intended outcomes.

In addition, we explored the distinction between evidence of selection validity and evidence of selection effectiveness. Utility analysis is perhaps the most common evidence of effectiveness at the individual level of analysis. We also applied this distinction to the relationship between selection predictor scores aggregated to an organization level and organization-level outcomes where any causal linkage is ambiguous, corrupted, or obscured by the effects of other organizational factors. (See Chapter 5 in this volume for a description of the importance and the manner in which individual-level selection influences organization-level outcomes.) In this situation, the relationship between organization-level measures of predictor scores can often be the most important, ultimate objective for an organization's selection system, but this relationship does not have the same meaning as a validity relationship. One reason for addressing this distinction is to place some emphasis on this point to be clear that validity is not necessarily the only or even the most important objective for the selection professional.

This chapter demonstrates that, while validity is a unitary concept, the types of evidence supporting validity and the variety of design and implementation decisions that influence or are dependent on validity are hardly unitary.

NOTE

1. Throughout this chapter we use the term "scores" to generically refer to observed manifestations of a measurement procedure; thus, scores might be ratings, behavioral observations, test scores, etc.

REFERENCES

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 (joint committee). (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*, 201–238.

Americans with Disabilities Act, 42 U.S.C. § 12101 (1990) *et seq.*

ADA Amendments Act, 42 U.S.C. § 12101 (2008) *et seq.*

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., Doverspike, D., Munoz, G. J., Taylor, J. E., & Carr, A. E. (2014). The use of mobile devices in high-stakes, remotely delivered assessments and testing. *International Journal of Selection and Assessment, 22*, 113–123.

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.

Campbell, J. P. (2015). All general factors are not alike. *Industrial and Organizational Psychology, 8*(3), 428–434.

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

Campbell, W. J., & Reilly, M. E. (2000). Accommodations for persons with disabilities. In J. F. Kehoe (Ed.), *Managing selection in changing organizations* (pp. 319–367). San Francisco, CA: Jossey-Bass.

Carter, N. T., Dalal, D. K., Boyce, A. S., O'Connell, M. S., Kung, M., & Delgado, K. M. (2014). Uncovering curvilinear relationships between conscientiousness and job performance: How theoretically appropriate measurement makes an empirical difference. *Journal of Applied Psychology, 99*, 564–586. doi: 10.1037/a00334688

Connelly, B. S., & Ones, D. S. (2010). Another perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*, 1092–1122. http://doi.org/10.1037/a0021212

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 221–237). Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–300.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington DC: American Council on Education.

Davies, S. E., Connelly, B. L., Ones, D. S., & Birkland, A. S. (2015). The general factor in personality: The "Big One", a self-evaluative trait, or a methodological gnat that won't go away? *Personality and Individual Differences, 81*, 13–22.

Delany, T., & Pass, J. (2005). *Design and validation of an unproctored cognitive ability tests*. Paper presented at the 20th annual conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

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* (pp. 29–81). San Francisco, CA: Jossey-Bass.

Green, J. P., Bradshaw, P., Kelly, E. D., Zhu, M., Dalal, R. S., & Meyer, R. D. (2015). *Personality strength: Operationalization and relationship with within-person performance variation*. Paper presented at the 30th annual conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA.

Gonzalez-Mule, E., Mount, M. K., & Oh, I.-S. (2014). A meta-analysis of the relationship between general mental ability and nontask performance. *Journal of Applied Psychology, 99*, 1222–1243. doi: 10.1037/a0037547

Guion, R. M. (1974). Open a new window: Validities and values in psychological measurement. *American Psychologist, 29*, 287–296.

Guttman, A. (2012). Legal constraints on personnel selection decisions. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 686–720). Oxford, UK: The Oxford University Press.

Hausknecht, J. P. (2010). Candidate persistence and personality test practice effects: Implications for staffing system management. *Personnel Psychology, 63*, 299–324. doi: 10.1111/j.1744–6570.2010.01171.x

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373–385. doi: 10.1037/0021–9010.92.2.373

Hoffman, C. C., Rashovsky, B., & D'Egidio, E. (2007). Job component validity: Background, current research, and applications. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 82–121). San Francisco, CA: Jossey-Bass.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking and employment selection. *Journal of Applied Psychology, 92*, 1270–1285. doi: 10.1037/0021–9010.92.5.1270

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75*, 581–595.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29*, 340–362.

Illingworth, A. J., Morelli, N. A., Scott, J. C., & Boyd, S. L. (2015). Internet-based, unproctored assessments on mobile and non-mobile devices: Usage, measurement equivalence, and outcomes. *Journal of Business and Psychology, 30*, 325–343.

International Test Commission. (2006). International guidelines on computer-based testing and Internet-delivered testing. *International Journal of Testing, 6*, 143–172.

Johnson, J. W. (2007). Synthetic validity: A technique of use (finally). In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence*. (pp. 122–158). San Francisco, CA: Jossey-Bass.

Kaminsky, K. A., Hemingway, M. A. (2009). To proctor or not to proctor: Balancing business needs with validity in online assessment. *Industrial and Organizational Psychology, 2*, 24–26.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kehoe, J. F., Brown, S., & Hoffman, C. (2012). The life cycle of successful selection programs. In N. Schmitt (Ed.), *The Oxford handbook of personnel selection and assessment* (pp. 903–938). Oxford, UK: The Oxford University Press.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*, 179–188.

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

Le, H., Oh, I., Robbins, S. B., Remus, I., Holland, E., & Westrick, P. (2011). Too much of a good thing: Curvilinear relationships between personality traits and job performance. *Journal of Applied Psychology, 96*, 113–133. doi: 10.1037/a0021016

Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology, 84*, 817–824.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58*, 981–1007. doi: 10/1111/j.1744–6570.2005.00713.x

Lievens, F., Buyse, T., & Sackett, P. R., & Connelly, B. S. (2012). The effects of coaching on situational judgment tests in high-stakes selection. *International Journal of Selection and Assessment, 20*, 272–282. doi: 10.1111.j.1468–2389.2012.00599.x

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92*, 1672–1682.

Loevinger, J. (1957). Objective tests as instruments of psychological theory [Monograph No. 9]. *Psychological Reports, 3*, 635–694.

McCormick, D. J. (2001). *Lowering employee illness and rates of on-the-job accidents by screening for mental ability*. Paper presented at the 16th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

McPhail, S. M. (Ed.) (2007). *Alternative validation strategies: Developing new and leveraging existing validity evidence*. San Francisco, CA: John Wiley and Sons.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

Oh, I., Wang, G., & Mount, M. K. (2011). Validity of the observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology, 96*, 762–773. doi: 10.1037/a0021832

Ones, D. S., Vishwesvaran, C., & Reiss, A. D. (1996). The role of social desirability in personality testing for personnel decisions: The red herring. *Journal of Applied Psychology, 81*, 660–691.

Pearlman, K. (2009). Unproctored internet testing: Practical, legal and ethical concerns. *Industrial and Organizational Psychology, 2*, 14–19.

Powers, D. E. (1986). Relations of test item characteristics to test preparation / test practice effects: A quantitative summary. *Psychological Bulletin, 100*, 67–77.

Roe, R. A. (2005). The design of selection systems: Context, principles, issues. In A. Evers, N. Anderson, & O. Smit (Eds.), *Handbook of personnel selection* (pp. 73–97). Oxford, England: Blackwell.

Rotundo, M., & Sacket, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy capturing approach. *Journal of Applied Psychology, 87*, 66–80. doi: 10.1037/0021–9010.87.1.66

Sackett, P. R., Burris, L. R., & Ryan, A. M. (1989). Coaching and practice effects in personnel selection. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 145–183). Oxford, England: John Wiley & Sons.

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.

Schmitt, N., & Landy, F. J. (1993). The concept of validity. In N. Schmitt, W. C. Borman, & Associates (Eds.), *Personnel selection in organizations* (pp. 275–309). San Francisco, CA: Jossey-Bass.

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

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organization Research Methods, 15*, 463–487. doi: 10.1177/1094428112444611

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25–39. doi: 10.1037/0021–9010.91.1.25

Tippins, N. (2002). Issues in implementing large-scale selection programs. In J. W. Hedge & E. D. Pulakos (Eds.), *Implementing organization interventions: Steps, processes, and best practices* (pp. 232–269). San Francisco, CA: Jossey-Bass.

Tippins, N. (2012). Implementation issues in employee selection testing. In N. Schmitt (Ed.), *The Oxford handbook of personnel selection and assessment* (pp. 881–902). Oxford, UK: The Oxford University Press.

Van Iddekinge, C. H., Morgeson, F. P., Schleicher, D. J., & Campion, M. A. (2011). Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. *Journal of Applied Psychology, 96*, 941–955. doi: 10.1037/a0023562

Vecchione, M., Alessandri, G., & Barbaranelli, C. (2012). Paper-and-pencil and web-based testing: The measurement invariance of the Big Five tests in applied settings. *Assessment, 19*, 243–246. doi: 10.1177/1073.191111419091

Villado, A. J., Randle, J. G., & Zimmer, C. U. (2016). The effect of method characteristics on retest score gains and criterion-related validity. *Journal of Business and Psychology, 31*, 1–16. doi: 10.1007/s.10869–015–9408–7

Walmsley, P. T., & Sackett, P. R. (2013). Factors affecting potential personality retest improvement after initial failure. *Human Performance, 26*, 390–408. doi: 10.1080/08959285.2013.836196

Weiner, J. A., & Morrison, J. D. (2009). Unproctored online testing: Environmental conditions and validity. *Industrial and Organizational Psychology, 2*, 27–30.