Overview of the Validation Process

The essential principle in the evaluation of any selection procedure is that evidence be accumulated to support an inference of job relatedness. Selection procedures are demonstrated to be job related when evidence supports the accuracy of inferences made from scores on, or evaluations derived from, those procedures with regard to some important aspect of work behavior (e.g., quality or quantity of job performance, performance in training, advancement, tenure, termination, or other organizationally pertinent behavior). Although this document focuses on individual performance, group and organizational performance may also be relevant criteria.

Any claim of validity made for a selection procedure should be documented with appropriate research evidence built on the principles discussed in this document. Promotional literature or testimonial statements should not be used as evidence of validity.

The *Principles* embraces the *Standards'* definition of validity as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (AERA et al., 1999, p. 184). Validity is the most important consideration in developing and evaluating selection procedures. Because validation involves the accumulation of evidence to provide a sound scientific basis for the proposed score interpretations, it is the interpretations of these scores required by the proposed uses that are evaluated, not the selection procedure itself.

The *Standards* notes that validation begins with "an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation refers to the constructs or concepts the test is intended to measure" (AERA et al., 1999, p. 9). Examples of such constructs or concepts include arithmetic proficiency, managerial performance, ability to design a Web page, oral presentation skills, conscientiousness, and ability to trouble-shoot technical problems with equipment on an assembly line. A clear description of the construct or conceptual framework that delineates the knowledge, skills, abilities, processes, and characteristics to be assessed should be developed.

In the early 1950s, three different aspects of test validity were discussed—content, criterion related, and construct. Since that time, the conceptualization of validity evidence has undergone some modification, moving from three separate aspects of validity evidence to the current *Standards'* view of validity as a unitary concept with different sources of evidence contributing to an understanding of the inferences that can be drawn from a selection procedure. Nearly all information about a selection procedure, and inferences about the resulting scores, contributes to an understanding of its validity. Evidence concerning content relevance, criterion relatedness, and construct meaning is subsumed within this definition of validity.

The validity of any inference can be determined through a variety of different strategies for gathering evidence. The *Standards* notes that while different strategies for gathering evidence may be used, the primary inference in employment contexts is that a score on a selection procedure predicts subsequent work behavior. Even when the validation strategy used does not involve empirical predictor-criterion linkages, such as when a user relies on test content to provide validation evidence, there is still an implied link between the test score and a criterion. Therefore, even when different strategies are employed for gathering validation evidence, the inference to be supported is that scores on a selection procedure can be used to predict subsequent work behavior or outcomes. Professional judgment should guide the decisions regarding the sources of evidence that can best support the intended interpretation and use.

The quality of validation evidence is of primary importance. In addition, where contradictory evidence exists, comparisons of the weight of evidence supporting specific inferences to the weight of evidence opposing such inferences are desirable.

The *Standards* discusses five sources of evidence that can be used in evaluating a proposed interpretation of selection procedure test scores for a particular use: (a) relationships between predictor scores and other variables, such as test-criterion relationships, (b) content, (c) internal structure of the test, (d) response processes, and (e) consequences of testing. Given that validity is a unitary concept, such categorizations refer to various sources of evidence rather than distinct types of validity. It is not the case that each of these five sources is an alternative approach to establishing job relatedness. Rather, each provides information that may be highly relevant to some proposed interpretations of scores, and less relevant, or even irrelevant to others.

Sources of Evidence

Evidence Based on the Relationship Between Scores on Predictors and Other Variables

This form of evidence is based on the empirical relationship of predictor scores to external variables. Two general strategies for assembling empirical evidence apply. The first strategy involves examining the relationship between scores on two or more selection procedures measuring the same construct hypothesized to underlie the predictor measure. Evidence that two measures are highly related and consistent with the underlying construct can provide convergent evidence in support of the proposed interpretation of test scores as representing a candidate's standing on the construct of interest. Similarly, evidence that test scores relate differently to other distinct constructs can contribute to discriminant evidence of validity. Note that convergent and discriminant evidence does not in and of itself establish job

relatedness, which requires evidence linking selection procedure scores to work-relevant behavior.

A second strategy typically involves relating a test or other selection procedure to a criterion. This strategy has historically encompassed two study designs: predictive and concurrent. A predictive study examines how accurately test scores predict future performance. In a concurrent study, predictor and criterion data are collected during a relatively simultaneous time frame although the objective remains to predict performance.

Content-Related Evidence

Test content includes the questions, tasks, format, and wording of questions, response formats, and guidelines regarding administration and scoring of the test. Evidence based on test content may include logical or empirical analyses that compare the adequacy of the match between test content and work content, worker requirements, or outcomes of the job.

Evidence Based on the Internal Structure of the Test

Studies that examine the internal structure of a test and the relationship among its items or tasks (e.g., work samples) can provide additional evidence of how test scores relate to specific aspects of the construct to be measured. Such evidence typically includes information concerning the relationships among items and the degree to which they represent the appropriate construct or content domain. For example, evidence of a high degree of item homogeneity is appropriate when a single dimension or singular construct is to be measured, but if the conceptual framework requires a more complex structure, overall consistency among items may not provide appropriate evidence of the internal structure of the test. When a multidimensional factor structure is proposed, evidence supporting inferences concerning the validity of score interpretations for the subcomponents in the predictor may be appropriate.

Evidence Based on Response Processes

In employment contexts, evidence based on response processes is necessary when claims are made that scores can be interpreted as reflecting a particular response process on the part of the examinee. For example, if a claim is made that a work sample measures use of proper techniques for resolving customer service problems, then simply assessing whether the problem is resolved is not enough. Evidence based on both cognitive and physical response processes may provide additional evidence of validity. Examining the processes used by individuals in responding to performance tasks or test questions can provide such evidence. Often evidence regarding individual responses can be gathered by (a) questioning test takers about their response strategies, (b) analyzing examinee response times on computerized assess-

ments, or (c) conducting experimental studies where the response set is manipulated. Observations of how individuals engage in performance tasks can also illustrate the extent to which the task is eliciting behavior related to the intended construct as opposed to behavior more related to irrelevant constructs. However, in many employment contexts such evidence is irrelevant to the proposed use, as is the case where the only claim made is that the scores on the selection procedure are predictive of a particular work outcome.

Evidence Based on Consequences of Personnel Decisions

In recent years, one school of thought has advocated incorporating examination of consequences of the use of predictors in the determination of validity. This perspective views unintended negative consequences as weakening the validity argument. Although evidence of negative consequences may influence policy or practice decisions concerning the use of predictors, these *Principles* and the *Standards* take the view that such evidence is relevant to inferences about validity only if the negative consequences can be attributed to the measurement properties of the selection procedure itself.

Subgroup differences resulting from the use of selection procedures are often viewed as a negative consequence of employment selection. Group differences in predictor scores and selection rates are relevant to an organization and its employment decisions, yet such differences alone do not detract from the validity of the intended test interpretations. If the group difference can be traced to a source of bias or contamination in the test, then the negative consequences do threaten the validity of the interpretations. Alternatively, if the group difference on the selection procedure is consistent with differences between the groups in the work behavior or performance predicted by the procedure, the finding of group differences could actually support the validity argument. In this case, negative consequences from test use constitute a policy issue for the user, rather than indicate negative evidence concerning the validity of the selection procedure.

A different example of negative consequences is also helpful. An organization that introduces an integrity test to screen applicants may assume that this selection procedure provides an adequate safeguard against employee theft and will discontinue 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 decisions subsequent to the introduction of the test may have had an unanticipated, negative consequence on the organization. Such consequences may lead to policy or practice decisions to reduce the negative impact. However, such consequences do not threaten the validity of inferences that can be drawn from the integrity tests, as the consequences are not a function of the test itself.

Planning the Validation Effort

Before a validation effort is planned, the proposed uses of the selection procedures being considered must be based on an understanding of the work performed, and the needs and rights of the organization and its present and prospective employees. These proposed uses should be consistent with professional, ethical, and legal responsibilities. Validation begins with a clear statement of the proposed uses as well as the intended interpretations and outcomes and should be designed to determine how well the proposed uses will be achieved.

Selection procedures used in the overall selection process should be supported by validity evidence. When a selection decision is based on multiple components combined into a composite, evidence for the final decision has primary importance. The validation effort should accumulate evidence that generalizes to the selection procedure and work behavior in the operational setting. The design of this effort may take many forms such as single local studies, consortium studies, meta-analyses, transportability studies, or synthetic validity/job component studies. More than one source of evidence or validation strategy may be valuable in any one validation effort.

In planning a validation effort for personnel decisions, three sources of evidence are most likely to be relevant: relationships to measures of other variables, content-related evidence, and internal structure evidence. Under some circumstances, evidence based on response processes and evidence based on consequences may be important to consider. The decision to pursue one or more of these sources of evidence is based on many considerations including proposed uses, types of desired selection procedures, availability and relevance of existing information and resources, and strength and relevance of an existing professional knowledge base. Where the proposed uses rely on complex, novel, or unique conclusions, multiple lines of converging evidence may be important.

The design of the validation effort is the result of professional judgment balancing considerations that affect the strength of the intended validity inference with practical limitations. Important considerations include (a) existing evidence, (b) design features required by the proposed uses, (c) design features necessary to satisfy the general requirements of sound inference, and (d) feasibility of particular design features.

Existing Evidence

An important consideration in many validation efforts is whether sufficient validity evidence already exists to support the proposed uses. The availability and relevance of existing evidence and the potential information value of new evidence should be carefully weighed in designing the validation effort. All validity conclusions are generalizations from the results in the validation setting to selection procedures and work behavior in the oper-

ational setting. The information value of existing and possible new evidence is based on the many factors that affect the strength of this generalization.

Existing evidence provides information value where it establishes statistical dependability and supports the generalization from the validation setting(s) to the operational settings. Where such evidence has been accumulated, it may provide a sufficient rationale for inferring validity in the operational setting and may support a decision not to gather additional evidence. Such inferences depend on evidence of validity rather than mere claims of validity. Advances in meta-analysis methods and a growing knowledge base of meta-analysis results have established considerable validation evidence for cognitive ability measures, and evidence is accruing for noncognitive measures such as personality and physical abilities. However, existing evidence alone may not be sufficient to support inferences of validity in a given situation.

Validity conclusions based on existing evidence may be strengthened by evidence from more than one method especially where the validity inference depends heavily on some underlying or theoretical explanatory concept or construct. In such cases, different methods may not support the same conclusions about the underlying explanatory concepts or constructs. For example, factor analyses of test scores may not replicate factor analyses of ratings of the same attributes. In these situations, convergent and discriminant evidence across multiple methods may be important.

Proposed Uses

In designing a validation effort, whether based on existing evidence, new evidence, or both, primary consideration should be given to the design features necessary to support the proposed uses. Examples of such features include the work to be targeted (e.g., one job title or a family of related work), the relevant candidate pool (e.g., experienced or nonexperienced candidates), the uniqueness of the operational setting (e.g., one homogeneous organization or many different organizations), and relevant criterion measures (e.g., productivity or turnover).

Requirements of Sound Inference

Primary consideration should also be given to the general requirements of sound validity inferences including measurement reliability and validity, representative samples, appropriate analysis techniques, and controls over plausible confounding factors. People who provide information in the validation effort should be knowledgeable and qualified for the tasks they are asked to perform and content they are asked to contribute.

Feasibility

Validation planning must consider the feasibility of the design require-ments necessary to support an inference of validity. Validation efforts may be limited by time, resource availability, sample size, or other organization con-straints including cost. In some situations these limits may narrow the scope of appropriate generalizations, but in other situations they may cause design flaws leading to inaccurate generalizations. While validation efforts with a narrow focus may have value, poorly executed validation efforts may lead the employer to reject beneficial selection procedures or accept invalid ones. Misleading, poorly designed validation efforts should not be undertaken.

Analysis of Work

Historically, selection procedures were developed for specific jobs or job families. This often remains the case today, and traditional job analysis methods are still relevant and appropriate in those situations. However, organizations that experience rapid changes in the external environment, the nature of work, or processes for accomplishing work may find that tradi-tional jobs no longer exist. In such cases, considering the competencies or broad requirements for a wider range or type of work activity may be more appropriate. Competency models are often used by organizations for many different purposes (Schippmann et al., 2000). When they are intended to support the underlying validity or use of a selection procedure, these *Prin-ciples* apply. The term "analysis of work" is used throughout this document and subsumes information that traditionally has been collected through job analysis methods as well as other information about the work, worker, organization, and work environment. The focus for conducting an analysis of work may include different dimensions or characteristics of work includ-ing work complexity, work environment, work context, work tasks, behav-iors and activities performed, or worker requirements (e.g., knowledge, skills, abilities, and other personal characteristics [KSAOs]).

Purposes for Conducting an Analysis of Work

There are two major purposes for conducting an analysis of work. One purpose is to develop selection procedures. Part of this process is an analy-sis of work that identifies worker requirements including a description of the general level of ability, skill, knowledge, or other characteristics needed. Such an analysis of work would determine the characteristics workers need to be successful in a specific work setting, or the degree to which the work requirements are similar to requirements for work performed elsewhere. The other purpose is to develop or identify criterion measures by assembling the information needed to understand the work performed, the setting in which the work is accomplished, and the organization's goals.

There is no single approach that is the preferred method for the analysis of work. The analyses used in a specific study of work are a function of the nature of work, current information about the work, the organizational setting, the workers themselves, and the purpose of the study. Understanding the organization's requirements or objectives is important when selecting an appropriate method for conducting an analysis of work. The choice of method and the identification of the information to be gathered by that method should include the relevant research literature.

Level of Detail

The level of detail required of an analysis of work is directly related to its intended use and the availability of information about the work. A less detailed analysis may be sufficient when there is already information descriptive of the work. A less detailed analysis may be appropriate when prior research about the job requirements allows the generation of sound hypotheses concerning the predictors or criteria across job families or organizations. When a detailed analysis of work is not required, the researcher should compile reasonable evidence establishing that the job(s) in question are similar in terms of work behavior and/or required knowledge, skills, abilities, and/or other characteristics, or falls into a group of jobs for which validity can be generalized. Situations that require a more detailed analysis of work may include those in which there is little existing work information available and the organization intends to develop predictors of specific job knowledge.

Any methods used to obtain information about work or workers should have reasonable psychometric characteristics and should be understood by the participants. Lack of consensus about the information contained in the analysis of work should be noted and considered further. Current job descriptions or other documents may or may not serve the immediate research purpose. Such information needs to be evaluated to determine its relevance and usefulness.

In some instances, an analysis of work may be the basis for assigning individuals to or selecting individuals for future jobs that do not exist at present. In other instances, an analysis of work may be used for transitioning workers from current to future work behaviors and activities. In both instances, the future work behaviors and activities, as well as the worker requirements may differ markedly from those that exist at present. Similarly, the work environment in which an organization operates also may change over time. For example, technology has permitted many individuals to work from virtual offices and replaced many functions that were previously conducted by individuals. Further, the global environment has expanded geographical boundaries and markets for many organizations. Procedures similar to those used to analyze current work requirements may be applicable for conducting an analysis of work in environments of rapid

change. However, other approaches that may be more responsive to the complexities of the emerging work environments are more appropriate (Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1999; Schneider & Konz, 1989). The central point in such instances is the need to obtain reliable and relevant job information that addresses anticipated behaviors, activities, or KSAOs.

If there is reason to question whether people with similar job titles or work families are doing similar work, or if there is a problem of grouping jobs with similar complexity, attributes, behaviors, activities, or worker KSAOs, inclusion of multiple perspectives and incumbents in an analysis of work may be necessary. Even when incumbents are in positions with similar job titles or work families, studying multiple incumbents may be necessary to understand differences in work complexity, work context, work environment, job behaviors, or worker KSAOs as a function of shift, location, variations in how work is performed, and other factors that may create differences in similar job titles or worker families.