Principles for the Validation and
Use of Personnel Selection Procedures

Statement of Purpose

This statement of principles has been adopted by the Executive Committee of the Society for Industrial and Organizational Psychology (Division 14 of the American Psychological Association) as the official statement of the Society concerning procedures for conducting validation research and personnel selection. Its purpose is to specify principles of good practice in the choice, development, evaluation, and use of personnel selection procedures.

Such selection procedures include, but are not limited to, standardized paper-and-pencil tests, performance tests, work samples, personality inventories, interest inventories, projective techniques, tests of honesty or integrity including polygraph examinations, assessment center evaluations, biographical data forms or scored application blanks, interviews, educational requirements, experience requirements, reference checks, physical requirements such as height or weight, physical ability tests, appraisals of job performance, computer-based test interpretations, estimates of advancement potential, or any other selection instrument, whenever any one or a combination of them is used or assists in making a personnel decision. Personnel decisions are decisions to hire, promote, terminate, transfer, and/or take other actions that affect one's employment status.

Selection procedures are used with the implicit assumption that some important aspect of behavior on the job (including performance in training, advancement, termination, or other organizationally pertinent behavior as well as quality or quantity of job performance) can be predicted from scores on the selection procedure. The essential principle in the evaluation of any selection procedure is that evidence be accumulated to support an inference of job relatedness.

This document is a revision of the second edition of the Principles for the Validation and Use of Personnel Selection Procedures published in 1980. The revision is stimulated by theoretical and research developments since the previous edition and the publication of Standards for Educational and Psychological Testing in 1985 by the American Educational Research Association, American Psychological Association and the National Council on Measurement in Education. The purposes of the revision are to bring the Principles up to date scientifically, to make them consistent with the Standards, and to reduce possible ambiguities regarding good practice in the use of selection procedures in making employment decisions.
The Principles provide:
(1) principles upon which the conduct of selection research may be based,
(2) principles for application and use of valid selection procedures, and
(3) information that may be helpful to personnel managers and others responsible for authorizing or implementing validation efforts.

The interests of some people will not be addressed in the Principles since they are not intended:
(1) to be a technical translation of existing or anticipated regulation,
(2) to be a substitute for adequate training in validation procedures,
(3) to be exhaustive (although they cover the major aspects of validation), or
(4) to freeze the field to prescribed practices and so limit creative endeavors.

The last point deserves emphasis. Traditional technology calls for a showing that (a) assessments made by a particular method (or combination of methods) are useful for predicting behavior in some aspect of employment, and (b) that the predictions can be made within an acceptable allowance for error (usually expressed in terms of coefficients of correlation or percentage of misclassifications). The use here of "predicting" and "predictions" implies no preference for a criterion-related procedure which uses a predictive strategy. While all assessments made by a selection procedure are secured with the express or implied expectation that they are related to one or more important aspects of job behavior, it is possible to support these assessments with other than a predictive criterion-related procedure. Acceptable procedures other than those discussed in this Principles may be developed in the future.

The Principles are intended to be consistent with the Standards (AERA et al., 1985). The Standards primarily address psychometric issues while the Principles primarily address the problems of making decisions in employee selection, placement, promotion, etc. The major concern of the Standards is general; the primary concern of the Principles is that performance on a test (or other basis for decision) is related to performance on a job or other measures of job success. Like the Standards, the Principles stated here present ideals toward which the members of this Society and other researchers and practitioners are expected to strive. Circumstances in any individual study or application affect the importance of any given principle. It is neither likely nor expected that anyone can satisfy completely the ideal of every applicable principle.

The likelihood that some principles cannot be completely met raises the question of relative levels of stringency in applying them to individual circumstances. The importance of a principle depends primarily on the consequences of failure to satisfy it. In selection research, where failure to adhere to a given principle would create a serious possibility of an erroneous decision about the validity or job-relatedness of a selection procedure, it is particularly important to adhere to proper procedures. Where few or no such consequences ensue, the principle in question assumes reduced importance. Similarly, in making personnel decisions based on any selection procedure, the importance of adherence to the Principles depends on the consequences of error which may result in physical, psychological, or economic injury to people or reduce the safety or operating efficiency of the organization. If such consequences are likely, principles should be followed more rigorously than in less crucial situations.

It is important to repeat here the following statement from the 1974 edition of the Standards, cited in full in the first and second edition of the Principles:

A final caveat is necessary in view of the prominence of testing issues in litigation. This document is prepared as a technical guide for those within the sponsoring professions; it is not written as law. What is intended is a set of standards to be used, in part, for self-evaluation by test developers and test users. An evaluation of their competence does not rest on the literal satisfaction of every relevant provision of this document. The individual standards are statements of ideals or goals, some having priority over others. Instead, an evaluation of competence depends on the degree to which the intent of this document has been satisfied by the test developer or user (AERA et al., 1974, p. 8).

And, from the 1985 edition of the Standards, . . . it should be emphasized, however, that in legal proceedings and elsewhere professional judgment based on the accepted corpus of knowledge always plays an essential role in determining the relevance of particular standards in particular situations (AERA et al., 1985, p. 2). The Principles are also intended to represent the consensus of professional knowledge and thought as it exists today, but not a consensus omnium since one is unattainable. Also, it is to be noted that personnel selection research and development is an evolving field in which techniques and decision-making models are subject to change. The Principles contain references that provide support for the principles enunciated and are recommended for further reading. Both researchers and practitioners are expected to maintain an appropriate level of awareness of developments in the field.

This revision, as with its earlier versions, is concerned primarily with personnel selection procedures. This is not to imply that other areas of personnel administration are exempt from meeting the standards of good professional practice. For example, performance measures may be used for training and career development; the Standards refer to such use. Those sections of the Principles which refer to the development of criterion measures are relevant in such instances.
Definition of Validity

Any claim made for any selection procedure should be documented with appropriate research evidence built on a foundation of systematic procedures like those discussed herein. Promotional literature or testimonial statements may not be used as evidence of the validity of a procedure.

The Standards discuss validity in these terms:

"Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself" (AERA et al., 1985, p. 9).

The specific kind of inference to be made from scores obtained by a selection procedure is ultimately an inference about probable job behavior. The specific inference investigated by the researcher should be clearly stated. The validity of any inference can be determined in a variety of ways. Although more than one line of evidence may be desirable in some instances, for inferences from some test scores a single line of evidence may be appropriate and sufficient. The quality of evidence is of primary importance, and a single line of solid evidence is preferable to numerous lines of evidence of questionable quality.

The Principles discuss three validation strategies—content, criterion-related, and construct—separately only to acknowledge traditional presentations and avoid an abrupt departure from tradition. The reader is advised that in concept, and many times in methodology, the three cannot logically be separated. The Principles use strategy instead of validity in labeling the three concepts to emphasize the interrelatedness of the three approaches to collecting evidence. The Principles also contain discussions of the generality vs. specificity issue in validation. The need to develop selection procedures with generality is emphasized, not only to acknowledge practical considerations, but also to further the search for understanding of selection measures.

Objectives of Validation Efforts

Before a selection procedure is considered, or a validation effort is planned, the researcher should be able to make a clear statement of the objective of the procedure. The statement of purpose must be based on an understanding of the work performed on the job and of the needs and rights of the organization and its present and prospective employees. There should be clear objectives for the proposed selection procedure and the validation effort should be designed to determine how well they have been achieved. Objectives should be consistent with professional, ethical, and legal responsibilities.

All aspects of the decision-making process should make a valid contribution to achievement of those objectives but of primary importance is the validity of the final selection decision.

Job Analysis

In conducting job analyses, the important consideration is understanding the organization's needs as they relate to the selection problem so that the researcher can formulate sound hypotheses about relationships among predictors and criteria. A systematic job analysis should provide the investigator with information to develop relevant criteria and make good judgments about predictors. However, a less detailed job analysis may be all that is required because there is so much previous job analytic work on the occupation of interest or because past research on the job allows the generation of sound hypotheses concerning predictors and criteria can be developed with little reference to a specific job analysis in a particular organization. When a systematic new job analysis is not completed, the researcher should compile reasonable evidence which establishes that the jobs in question are similar in terms of work behavior and/or required knowledge, skills, and abilities.

Job analysis is essential to the development of a content-oriented procedure or to the justification of a construct considered important to job behavior. Several job analysis procedures exist, each differing in its possible contribution to the objectives of a validity study. Several recent books provide summaries of various job analysis procedures and discussions of their relative utility in various situations (Gael, 1983; Levine, 1983; McCormick, 1979). Whatever job analysis method is used, it should be used carefully. Sources of job information should be credible. Any scales used to evaluate tasks and knowledges, skills, and abilities (KSAs) should have reasonable psychometric characteristics. Lack of consensus among job experts regarding tasks and KSAs should be noted and carefully considered. Current job descriptions or other documents may or may not serve the immediate research purpose.

One purpose of the job analysis is to assemble the information needed to understand the work performed on the job and the setting in which the work is accomplished. The written documentation of the results of the job analysis usually should include a job description. The job description
should communicate what job incumbents do, the resources and tools they use in performing their jobs, and any unusual or extreme characteristics of the job setting. The job description should identify the most important and complex of the job duties or activities.

Another purpose of the job analysis is the identification of worker specifications which include a description of the general level of skill or knowledge required to perform the job duties. Inferences about the worker characteristics are made by the researcher or job experts by combining knowledge of the work performed and what workers have to do to perform the work.

The amount of information required depends on the purposes to be served by the selection procedure and on the nature of the problems that the selection procedure is to address. It may be necessary to study more incumbents if there is reason to question whether people with similar job titles are, in fact, doing similar work, or if there is a problem of grouping jobs with similar tasks or responsibilities than if the jobs can clearly be placed in homogeneous groups.

In some situations, the major purpose of job analysis may be to provide information from which criterion measures may be developed; in others, the purpose may be to determine the characteristics candidates need to be successful. In any case, the choice of the job analysis method and the specific details of the method (e.g., number and kinds of questions to be asked, who is to answer them, the number and method of choice of incumbents to be observed or surveyed) depend on the objectives and on the constraints existing within the situation.

Factors to be considered include, among others, the nature of any personnel problems, the jobs, the situation, and the resources available. For example, the extent to which the objectives include assessing similarities among jobs or the formation of job families may be an important element in the choice of technique (Cornelius, Carron, & Collins, 1979; Pearlman, 1980).

**Criterion-Related Strategy**

Personnel selection procedures are used to predict future performance or other behavior. Evidence for criterion-related validity typically consists of a demonstration of a useful relationship between the selection procedure (predictor or predictors) and one or more measures of job relevant behavior (criterion or criteria). It is, therefore, vital that the choice of both predictors and criteria be made with great care.

A predictor is an aid to decision-making used in the context of selection, termination, placement, classification, promotion or other decision regarding personnel. Predictors include, but are not limited to, standardized ability tests, personality inventories, biographical data forms, situational tests, assessment center evaluations, ratings based on interviews, performance ratings, and evaluations of training or experience. (See Statement of Purpose, p. 1.) Standardized measures of a concept are preferred. A standardized measure is one for which standard directions and procedures for administration, scoring, and interpretation are used.

The principles discussed in this section, however, apply to all predictors.

**A. Determination of Feasibility.** A researcher contemplating a criterion-related validity study must first determine whether such a study is feasible. It is not always possible to conduct a well-designed or even an acceptable study. Although it may be argued that employers and the public are protected against the use of invalid tests because most errors reduce the estimated validity, it is also true that a poorly executed study may lead the employer to reject selection procedures that are beneficial to the organization and to the applicants. A poor study is not better than none.

Several considerations are relevant in determining feasibility. First, one must establish that the job is reasonably stable and not in a period of rapid evolution. Although validity coefficients seem to be robust across tasks and situations (Schmidt, Hunter, & Pearlman, 1981) for some predictor-criterion relationships, the traditional logic of validation research is that it is undertaken under conditions similar to those that are expected to exist when the results are made operational. If this assumption is not tenable, the researcher should either modify the validation strategy appropriately or postpone the study until the jobs and their settings are reasonably stable.

Second, it must be possible to obtain or to develop a relevant, reliable and uncontaminated criterion measure. Of these characteristics, the most important is relevance. A relevant criterion is one that reflects the relative standing of employees with respect to important job behavior. If such a criterion measure does not exist or cannot be developed, criterion-related validation is not feasible. It is inappropriate to do criterion-related studies using criteria which, although reliable and available, are either irrelevant or contaminated.

Third, a competent criterion-related validation should be based on a sample that is reasonably representative of the populations of people and jobs to which the results are to be generalized. Validities appear to be quite stable across both tasks and situations but there are influences, such as restriction of range in the predictor, the criterion, or both, that may distort an estimate obtained from a particular sample. When there is evidence of gross bias in sampling, the researcher either must estimate its impact and adjust for it, or must conclude that it is not feasible to conduct a criterion-related validation study.
Fourth, a criterion-related validity study that lacks adequate statistical power may leave the issue of validity unresolved. **Statistical power** refers to the probability of obtaining a statistically significant relationship between predictor and criterion in a sample if such a relationship exists in the population. Factors determining statistical power include the statistic computed, sample size, degree of range restriction in the predictor, reliability of the criterion, the probability of Type I error (the probability of concluding a significant relationship when none exists) and the size of the predictor-criterion relationship (Cohen, 1977). Combinations of these variables leading to low power can occur frequently in practice (Sackett & Wade, 1983; Schmidt, Hunter, & Urry, 1976). As a consequence, it is possible to conclude mistakenly that a significant predictor-criterion relationship does not exist when, in fact, it does. If the requirements cannot be met, the situation may not lend itself to a criterion-related validation effort.

**B. Design and Conduct of Criterion-Related Validity Studies.** If it has been determined that a criterion-related study is feasible, attention may then be directed to the design and conduct of a study.

A variety of designs can be identified. The traditional classification of predictive and concurrent validity is based on the presence or absence of a nontrivial time lapse between the collection of predictor and criterion data. In predictive designs, data on the selection procedure are collected at or about the time applicants are hired. After employees' performance levels have stabilized caution data are collected. In concurrent designs, the predictor and criterion data are collected on job incumbents at approximately the same time.

There are, however, other differences. Designs may differ in the time of predictor data collection relative to a hiring decision or the time at which employees start to work—before, simultaneously, shortly after, or after a substantial time period on the job. Designs may differ with respect to the basis for the hiring decision: the predictor itself, an "existing" predictor, a random procedure, or some combination of these. They may differ with respect to the population sampled in collecting predictor data. For example, the design may use an applicant population or a population of newly-hired employees not yet on the job, recent employees not yet fully trained, or employees with the full range of individual differences in experience.

For tests of cognitive abilities, estimates of validity obtained from predictive versus concurrent designs may be expected to be comparable (Bemis, 1968; Pearlman, Schmidt, & Hunter, 1980; Barrett, Phillips, & Alexander, 1981). Meta-analyses by Schmitt, Gooding, Noe, and Kirsch (1984) found only small differences in the estimates of population validity differences for three classes of designs. This finding cannot automatically be generalized to all situations and to other types of predictors and criteria. For example, concurrent validities of paper-and-pencil measures of personality traits may not provide accurate estimates of predictive validity. Other design differences may be important for other reasons. Guion and Cranny (1982) described five different concurrent designs that differed substantially in the data they provide for making population validity estimates based on corrections for restriction of range. Sussman and Robertson (1986) identified eleven designs which varied in susceptibility to threats to the Cook and Campbell (1979) criteria for research design.

In planning a validation study, the researcher should identify the design characteristics (nature of appropriate samples; need for, and feasibility of, specific statistical controls, corrections, or analyses; numbers of cases, etc.) required for a professionally acceptable validity study, determine how closely the design can approximate that ideal within situational constraints, and decide whether criterion-related validation is feasible.

1. **Criterion Development.** The researcher should obtain the necessary job information. In general, if criteria are chosen to represent job relevant activities or behaviors, the results of a formal job analysis are helpful in criterion construction. Although numerous procedures are available (see p. 4), there is not a clear choice of method. It is essential that information about the job be competently and systematically developed. If the goal of a given study is the prediction of nonperformance criteria such as tenure or absenteeism, a formal job analysis is not usually necessary, though an understanding of the job and its context is beneficial. Some considerations in criterion development follow.

a. **Criteria Should be Related to the Purposes of the Investigation.** Criteria should be chosen on the basis of relevance, freedom from contamination, and reliability rather than availability. This implies that the purposes of the research are (1) clearly stated, (2) acceptable in the social and legal context in which the organization functions, and (3) appropriate to the organization's needs and purposes. If adequate measures of important components of job performance are not attainable, the researcher should not substitute measures that are unrelated to the purposes of the study to achieve the appearance of broad coverage.

b. **All Criteria Should Represent Important Work Behaviors or Work Outputs, on the Job or in Job-Relevant Training, As Indicated By An Appropriate Review of Information About the Job.** Criteria need not be all-inclusive, but there should be
clear documentation of the reasoning behind the determination of the content of the criteria. Criteria need not be measures of actual job performance. In many cases, in fact, actual job performance measures may not possess the desirable characteristics specified above for criteria. Depending upon the job being studied and the purposes of the research, various criteria such as overall proficiency measured with a standard work sample, success in job relevant training, sales records, number of prospects called, turnover, or rate of advancement may be more appropriate (Wallace, 1965). When training scores are used as a criterion measure, it is necessary to insure their job relevance. The most commonly used measures of performance are ratings, usually by supervisors. The development of rating procedures should ordinarily be guided by job analyses if, for example, raters are expected to evaluate several different aspects of performance. It may also be necessary to train raters in the observation and evaluation of performance. Further, supervisors should be expected to be familiar enough with the demands of the job to evaluate overall performance.

c. The Possibility of Bias Should be Considered. Although a difference in the average criterion measures of different groups does not establish bias, such bias exists if a definable subgroup is rated consistently and spuriously high (or low) as compared to other groups. Conversely, if a group difference does exist but is not revealed by reliable ratings, there is bias. It is apparent that the presence or absence of bias cannot be detected from knowledge of criterion scores alone. There is no clear path to truth in these matters. A difference in criterion scores of older and younger employees, or day and night shifts could reflect bias in raters, equipment, or conditions, or it might reflect genuine differences in performance. The possibility of bias must be anticipated, and if it is found, it should be reduced or eliminated. The researcher should protect against bias insofar as is feasible, and use the best judgment possible in evaluating the data.

Efforts should be made to minimize contamination of criteria. A measure is contaminated to the extent that it includes unwanted systematic variance. Biasing factors may be considered contaminants; measurement techniques (e.g., complex verbal instructions for a psychomotor work sample) may introduce others. It is impossible to avoid (or even to know) completely all sources of contamination, but efforts should be made to minimize its effects. Measurement of some contaminating variables might enable the researcher to control statistically for them: in other cases, special diligence in the construction of the measurement procedure and in its use may be all that can be done.

d. If The Researcher Combines Several Criteria To Obtain A Single Variate, There Should Be A Rationale To Support The Rules of Combination. For example, it is often preferable to weight for relevance, although special circumstances sometimes argue otherwise. Ordinarily, it is better to assign unit or equal weights to the several criterion components than to attempt to develop precise empirical weights. At any rate, the rules of combination should be described.

When measures are combined, researchers should recognize that effective weights (i.e., the contributions of the various components to the variance of the composite) are unlikely to be the same as the nominal weights. Effective weights depend not only on the nominal weights assigned to the components but also to their standard deviations and intercorrelations. It is necessary to guard against circumstances, particularly in weighting measures of different lengths inversely according to their standard deviations, in which the least reliable components may have the greatest effective weight (Richardson, 1941).

e. It is Desirable, But Not Essential, That Criterion Measures be Highly Reliable. Reliability should be estimated by appropriate methods (e.g., Stanley, 1971). It should be recognized that criterion reliability places a ceiling on observed validity coefficients. Thus, the effect of criterion unreliability is to cause an underestimation of true validity.

2. Choice of Predictor. Many factors should influence choice of the predictor(s). Several of these follow.

a. Predictor Variables Should be Chosen for Which There is an Empirical, Logical, or Theoretical Foundation. This principle does not call for elegance in the reasoning underlying the choice of predictors so much as it does for having some reasoning. The rationale for a choice of predictors should be specified. A predictor is more likely to show validity if there is a good reason to suppose that a relationship exists between a predictor chosen and the behavior it is designed to predict. The research literature or the logic of development may provide the reason. This principle is not intended to rule out the application of serendipitous findings, but such findings, especially if based on small samples should be verified through replication in independent samples.
b. Preliminary Choices Among Predictors Should Be Based on the Researcher's Scientific Knowledge Without Regard for Personal Bias and Prejudice. The researcher's choice of trial predictors should be based on the findings of relevant research and resist the influence of personal interest, mere familiarity or expediency. On the other hand, the researcher must exercise critical judgment to achieve the parsimony in a predictor battery necessary to minimize predictor redundancy or the capitalization on chance which may occur with small samples.

c. In Measuring A Worker Specification, Predictors That Are More Objective Are to be Preferred. The assessment of candidates should be maximally dependent on their personal characteristics and minimally dependent on who made the assessment. Similarly, where non-test predictors like interviewer judgments are used, the researcher should develop procedures that will minimize error resulting from differences between judges.

d. Outcomes of Decision Strategies Should be Recognized as Predictors. Decision-makers who interpret and act upon predictor data interject something of themselves into the interpretive or decision-making process. The judgments or decisions thus become at least an additional predictor, or at the most, the only predictor. For example, if the decision strategy is to combine test and non-test data (reference checks, medical data, etc.) into a subjective judgment, the actual predictor is the judgment reached by the person who weights and summarizes all the information. It is this decision that should be validated in addition to the information which was available to the decision maker.

3. Choice of Sample. The generalizability of the research result depends in part on the sample. For example, several hundred subjects are not better than one hundred if the subjects in the larger sample are not relevant to the research purposes at hand.

a. The Sample for a Validation Study Should be Chosen Carefully. Whether the study is predictive or concurrent, a sample of incumbents is unlikely to be representative of the applicant group on all variables. Whether such characteristics as age, race, or sex affect predictor-criterion relationships is an empirical question, and the researcher should, therefore, rely on the research literature in making professional judgments about their possible relevance. Because many characteristics studied to date appear to have little or no effect on predictor-criterion relationships, no variable should be assumed to moderate validity coefficients in the absence of explicit evidence for such an effect. The research literature shows that validities across races (black vs. white) are usually comparable on cognitive selection tests (Linn, 1978).

b. The Sample Upon Which the Research is Based Should be Large Enough to Provide Adequate Statistical Power. A study that has only a low probability of detecting the true validity of the predictor provides little information. Statistical power may be increased to acceptable levels in a number of ways, the most obvious of which is to increase sample size by the addition of persons sampled from the same population.

c. An Extremely Large Sample or Replication is Required to Give Full Credence to Unusual Findings. Such findings include, but are not limited to, suppressor or moderator effects, nonlinear regression, benefits of configural scoring, or other potentially chance outcomes. Post hoc hypotheses in multivariate studies, and differential weightings of highly correlated variables are particularly suspect and should be replicated before they are accepted.

d. When Combining Data from Separate Samples, Both Jobs and Workers Should be Similar on Any Variables That May Affect Validity. If samples are comparable on these variables, pooling the samples would provide increased statistical power.

e. Dropouts and/or Exclusion of Outliers Should Be Explained. Occasionally, information on all variables for sample participants is not available or it was impossible to collect information on some participants. Information regarding the representativeness of the available sample should be provided perhaps including items such as their job level, tenure, ethnic status or sex. Exclusion of sample members because of extreme values on some variables should be noted and justified.

4. Procedural Considerations. The researcher should consider the probable use of any end products prior to the collection and analysis of data.

a. Validation Research Ordinarily Should be Directed to Entry Jobs, Immediate Promotions, or Jobs Likely to be Attained. Where a selection procedure is designed for a higher level job than that for which candidates are initially selected, that job may be considered an appropriate target job if the majority of the individuals who remain employed and available for advancement progress to the higher level within a reasonable period of time. Where a majority do not advance to the higher level job, it may still be acceptable to assess candidates for such job(s) if the validity study is conducted using criteria that reflect performance at the higher level along with criteria for
adequate performance at the entry level. The same logic may apply to situations in which people are rotated among jobs.

b. The Researcher May Consider AlternateCriterion-Related Research Methods That Offer a Sound Rationale. Examples include cooperative research on an industry-wide basis, consortia of small users, or gathering data for validity generalization. Such approaches generally require considerable effort in planning and in data analysis and considerable care in ascertaining the appropriateness of the data included from the different sources.

c. Procedures for Test Administration and Scoring in Validation Research Should Be Clearly Set Forth and Should Be Consistent with the Standardization Plan for Operational Use. Any specified operational characteristics (such as time limits, oral instructions, practice problems, answer sheets, and scoring formulas) should be clearly set forth and followed in validation research. Failure to do this essentially prohibits generalizations from the research to the operational context. The point of this principle is that for research to enhance the general body of knowledge, the critical research procedures must be consistent with those which are to be utilized in practice.

d. There Should Be At Least Presumptive Evidence For The Validity Of A Predictor Prior To Its Operational Use. If possible, predictors should be validated prior to operational use. Some researchers find this principle difficult to follow because of the need of the organization to make employment decisions. Where there is evidence from other studies or situations that makes valid prediction likely, it may be feasible to use the predictors immediately. The researcher should avoid situations that make it impossible or difficult to establish validity. For example, selection decisions based on an unvalidated test should not be so highly selective that severe restriction of range results. Data required for correction of restriction of range should be collected and maintained. If there is no firm basis for the presumption of validity, the researcher must judge whether the costs of postponing the use of the predictor are greater or less than the dangers of using it prematurely while collecting data.

e. Predictor Data and Criterion Measures Should Be Independent. If criterion ratings are collected from supervisors who know selection test scores, the two sets of data are not independent. The resulting validity coefficient depends on both the true relationship and the manipulation of ratings (consciously or unconsciously) to conform with scores. Such contamination should be avoided.

5. Data Analysis. The quality of the validity study depends as much on the appropriateness of the data analysis as on the data collected during the research. Researchers using computer facilities need to ascertain that the statistics used are appropriate. The importance of fully understanding statistics used is not reduced by the use of computers for calculation. Moreover, as with the choice of criterion or predictor variables, one should not choose a data analysis method simply because the computer package for it is readily available.

a. Data Analysis Should Yield Appropriate Information About the Relationship Between Predictor and Criterion Measures. The analysis should provide information about the magnitude and statistical significance of the relationships. Traditionally, a validity coefficient or similar statistic that has a probability of less than one in twenty of having occurred by chance if the true relationship is zero (Type I error) may be considered as establishing significant validity. Professional standards have never insisted on a specific level of significance. However, departures from this convention should be based on reasons that have been stated in advance (such as power functions, utility, and economic necessity).

Research on the power of criterion-related validation studies and research on validity generalization suggest that the dangers of Type I error may have been overemphasized. Researchers should give at least equal attention to the risks of Type II error; that is, the failure to detect validity where, with greater power in the analysis, significant correlations would be obtained. When multiple tests of significance are conducted for predictors with no cumulative literature indicating evidence of validity, the researcher should control for the possibility of observing some apparently significant relationships on a chance basis. When multivariate techniques are used, the number of cases should be large relative to the number of variables. The analysis should provide information about the strength of the relationship, usually a coefficient of correlation. Other methods (such as the slope of the regression line, expectancy tables, or the percentage of misclassifications) are acceptable and may be preferable in many situations. The analysis should also give information about the nature of the relationship and how it might be used in prediction. The information provided in the report should include number of cases, measures of central tendency and variability for both predictor and criterion variables, as well as the interrelationships among all variables studied.
b. Researchers Should Obtain as Unbiased an Estimate as Possible of The Validity of The Predictor in The Population in Which It Is Used. Observed validity coefficients are typically biased (Schmidt, Hunter, & Urry, 1976). Where range restriction operates to bias validity estimates, the appropriate adjustments should be made when possible. Adjustments for criterion unreliability should be made if an appropriate estimate of criterion reliability can be obtained. Researchers should make sure that reliability estimates used in making corrections are appropriate to avoid under or over-estimating validity. For example, in a criterion-related validity study in which the criteria are performance ratings, differences between raters and differences across time should be considered in estimating criterion reliability since internal consistency estimates, by themselves, may be inadequate. Both unadjusted and adjusted validity coefficients should be reported. Researchers should be aware that the usual tests of statistical significance do not apply to adjusted coefficients such as those adjusted for restriction of range and/or criterion unreliability. The adjusted coefficient is generally the best point estimate of the population validity. No adjustment of a validity coefficient for unreliability of the predictor should be reported unless it is clearly stated that the coefficient is theoretical and cannot be interpreted as reflecting the actual operational validity of the test.

c. Where Predictors Are To Be Used in Combination, Researchers Should Consider The Choice of The Mode of Combination. Researchers should be aware that nonlinear selection decision rules (e.g., random selection from among those scoring above a cutoff) typically reduce the validity and utility of selection procedures (Schmidt, Mack, & Hunter, 1984) when the true relationship is linear. When nonlinear selection rules are used, a clear rationale (e.g., in terms of administrative convenience or reduced testing costs) should be provided. Researchers should also consider the reliability of decisions based on nonlinear selection rules. Tests with linear relationships with job performance can be combined for use in either a linear manner (e.g., by summing scores on different tests) or in a nonlinear manner (e.g., by using multiple cutoffs) but the researcher should be aware of the productivity, administrative, legal and other implications of each choice.

d. Researchers Should Guard Against Overestimates of Validity Resulting from Capitalization on Chance. Especially when the research sample is small, estimates of the validity of a composite battery developed on the basis of a regression equa-

tion should be adjusted using the appropriate shrinkage formula (see Cattin, 1980) or be cross-validated on a new sample. The assignment of either rational or unit weights to predictors does not result in shrinkage in the usual sense. Where a smaller number of predictors is selected for use based on sample validity coefficients from a larger number included in the study, shrinkage formulas can be used only if the larger number is entered into the formula as the number of predictors, though this will produce a slightly conservative estimate of the cross-validated multiple correlation. Claudy (1969) presented a formula for cross-validated correlation which takes into account the total number of predictors considered as well as the number used in regression analyses.

e. The Results Obtained in Criterion-Related Validity Studies Should Be Interpreted Against the Background of the Relevant Research Literature. Cumulative research knowledge plays an important role in any science. Large bodies of research regarding relationships between test scores and job performance currently exist. A researcher may, after consulting the literature, conclude that the existing research base is sufficient to support the use of certain instruments without any additional criterion-related research. In fact, in some instances, particularly when the sample of research subjects is small or unrepresentative, an additional criterion-related study may lead to an erroneous conclusion about the validity of a selection instrument. In any event, the results of an individual validity study should be interpreted in light of the relevant research literature. A history of similar findings in the research literature lends additional credence to the results of individual studies so long as the history is truly representative of the experience of the field. On the other hand, dissimilar findings in a single study should be viewed with caution.

f. Ordinarily the Researcher Should Make an Assessment of the Practical Value of the Selection Procedure. The judgment of the practical value of a selection procedure is based on the evidence of validity, the selection ratio, the number to be selected, and the nature of the job. Expectancy tables can also be useful for this purpose, as can the Taylor-Russell tables. The literature on the impact of selection tests on the productivity of employees has provided estimates of utility that allow the use of regression equations (Brogden, 1949; Cronbach & Gieser, 1965; Schmidt, Hunter, McKenzie & Muldrow, 1979). Both projected productivity gains per employee and projected total productivity gains due to use of the selection procedures are
relevant in assessing their practical value (Boudreau & Berger, 1985; Cascio, 1982; Schmidt, Mack & Hunter, 1984). Utility analysis is a developing literature; wisdom suggests that all estimates of values to be placed in the equations should be chosen so as to lead to conservative estimates of utility. In addition to these conservative point estimates, minimal and maximal estimates should be presented when appropriate and needed.

g. Data Should be Free from Clerical Error. Data entry, coding, computational work, and reports should be checked carefully and thoroughly.

Differential Prediction

Fairness is a social rather than a psychometric concept. Its definition depends on what one considers to be fair. Fairness has no single meaning, and, therefore, no single statistical or psychometric definition. Fairness or lack of fairness is not a property of the selection procedure, but rather a joint function of the procedure, the job, the population, and how the scores derived from it are used.

Predictive bias is found when the mean criterion predictions for groups differentiated on some other basis than criterion performance are systematically too high or too low relative to mean criterion performance of the groups (Humphreys, 1952; Gordon, 1953; Cleary, 1968). Another term used to describe this phenomenon is differential prediction. Earlier use of this term to describe prediction in classification problems should not be confused with differential prediction applied to subgroup analyses. Although other definitions of bias have been introduced, only those based upon the regression model, as the definition given here is, have been found to be internally consistent (Petersen & Novick, 1976). Bias defined in this way has frequently been tested using an analysis of covariance procedure such as that provided by Gulliksen and Wilks (1950).

There is little evidence to suggest that there is differential prediction for the sexes, and the literature indicates that differential prediction on the basis of cognitive tests is not supported for the major ethnic groups (Schmidt, Pearlman, & Hunter, 1980; Hunter, Schmidt, & Rauschenberger, 1984). There is no compelling research literature or theory to suggest that cognitive tests should be used differently for different groups (National Academy of Sciences, 1982).

Content-Oriented Strategies

In content-oriented strategies, any inference about the usefulness of a score must be preceded by inferences based on the content and method of construction of the measurement instrument (Messick, 1975). For that reason, this section includes standards for the development or selection of the content-oriented instruments as well as inferences drawn from the test scores.

There is disagreement among professional personnel researchers in the use of the terms content-oriented strategy and content validity. Some people restrict these terms to situations in which a job domain is defined through job analysis by identifying important tasks, behaviors, or knowledge and the test (or criterion) is a representative sample of tasks, behaviors, or knowledge drawn from that domain. Others apply these terms to selection procedures in which more general worker specifications (such as general skills or abilities) are measured and match well those inferred from the job domain. In this section of the document we will emphasize the former narrower definition of content-oriented strategy. The second definition will be treated in the section on construct validity. Either approach may be an appropriate basis for the use of a selection procedure. There are also inconsistencies in the ways in which personnel researchers have used the terms knowledge, skills, and abilities. Researchers have frequently called the knowledge or skill related to a small group of tasks an ability. When ability is defined in this very specific way, content-oriented strategies may be sufficient. When referring to more general abilities such as reasoning or spatial ability, a construct-oriented strategy is likely to be necessary.

The procedure for the development or selection of a content-oriented predictor, if properly conducted, provides evidence that a selection procedure samples knowledge or skills required for job performance. The following provides guidance for the development or selection of predictors the use of which is based primarily on content.

Development or selection of a selection procedure on the basis of its relation to the content of the job requires that it include an appropriate sample of a relevant content domain. If a selection procedure is to be used for employment decisions, the relevant content domain is performance (or the knowledge or skill necessary for performance) on the job, in relevant job training, or on specified aspects of either (Lawshe, 1975). A procedure may sample a given job performance domain, but if that domain is not an important part of the job, the value of the procedure for employment purposes is negligible.

Content sampling is properly involved in the construction or choice of any selection procedure, the scores of which are to be interpreted as measures of achievement or as measures of work behavior. This discussion is limited, however, to situations in which the selection procedure is evaluated solely as a sample of job content. Content sampling is as useful in the construction and evaluation of criterion measures as it is for selection procedures used for employment decisions. This content sampling is not random but follows from the professional judgment of the
researcher. It may also involve the judgments of job experts and a job analysis that details critical tasks, important components of the work context, and behaviors required of the worker to perform the job. Procedures which systematically demonstrate that the content of the test actually reflects the content of the job strengthen claims of content validity.

A. The Job Content to be Sampled Should be Defined. The definition of the job content to be sampled should be based on an understanding of the job, organizational needs, labor markets, and other considerations which are relevant to personnel specifications and relevant to the organization's purposes. The domain need not include everything that is done on the job, but the investigator should describe the whole job as part of the job analysis, indicate what is included in the domain and explain why certain parts were or were not included. Different selection procedures may be validated in different ways and used to meet different purposes in selecting people for a given job. For example, it may be appropriate to represent one part of an overall job content domain with a procedure validated by criterion-related methods and a different part with a procedure developed by content sampling. The domains of interest in content sampling for selection are defined largely by the job activities that should be performed competently without further training. A job content domain may be further defined by specifying job components excluded from the domain.

In defining a content domain, it is essential that the degree of generality needed in a selection procedure be specified in advance. For example, the extent to which the job is likely to change should be known. If job changes are likely to be a problem, the researcher may wish to develop a selection procedure that is quite general; for example, eliminating material like specific sales prices that may change from month to month and concentrating on content that is less specific. The more a selection procedure has point-in-time fidelity to exact job operations, the less likely it is to have enough generality to remain appropriate when the job changes. Also, the more a selection procedure is a specific sample of a domain involved in one job, the less likely it is to apply to other similar jobs.

Specificity and generality form the ends of a continuum; the researcher must determine how general a selection procedure should be. The researcher should consider conditions that may affect the generality of the selection procedure. A selection procedure can be applied generally if there is a clear rationale based on the specific selection situation at hand, organizational needs, anticipated changes in technology, equipment, work assignments, and human and economic considerations. This principle also applies in the development of content-oriented criteria for use in criterion-related validity studies. The degree to which the results of criterion-related validity studies can be generalized depends partly on the generality of the criteria and their applicability over time and jobs.

B. Special Circumstances Should be Considered in Defining Job Content Domains. Domain definitions need not follow any prescribed format since there are many different ways to describe them depending on the job situation. It is necessary to assess possible measurement problems in advance of deciding to include a domain in the study. Generally, the more closely a test replicates a job behavior, the more difficult the procedure is to administer and score. For example, cleaning dirty mechanisms may be part of a mechanic's job, but it may be impossible to develop a work sample test so that every examinee would have the same amount and kind of dirt to remove. If that is the case, it would be appropriate to eliminate such cleaning tasks from the test domain. Seldom used symbols such as the hyphen or question mark appear in different places on different typewriter keyboards, so it might be appropriate to limit a typing domain to alpha and numeric characters that are standard on all typewriters. Also, a short course designed to select persons for a longer course should not be based on those portions of the longer course that depend on knowledge gained in the beginning lessons. In this situation, the domain should be defined only in terms of lessons that require no prior knowledge. Judgment must be used in defining a domain and the rationale involved must be explicitly described.

C. Job Content Domains Should be Defined on the Basis of Accurate and Thorough Information About the Job. Critical to the justification of a test on content bases is the similarity between the test content and the job content domain. Definition of a job content domain should be based on analysis of tasks, activities, or responsibilities of the job incumbents. Worker specifications are defined as the abilities, knowledge, job skills or other personal characteristics prerequisite to effective behavior in the domain. Determination of both requires a job analysis based on the judgments of informed persons such as production engineers, job incumbents, their supervisors, or personnel specialists.

When distinctions are made between abilities and skills, abilities are considered more general than skills. The use of the more general term does not imply that content validity is or is not a sufficient justification for the use of abilities or for such characteristics as empathy, dominance, leadership aptitude, general intelligence, and other broad psychological traits. Regardless of the label used to describe a set of test items, a content-oriented strategy requires a good match between test content and job content domains.

The validity of job requirements assessed by other than formal
tests may be established on the basis of their content. Requirements for or evaluation of personal history variables, such as specific prior training, experience, or achievement can be justified on the basis of the relationship between the content of the personal history experience and the content of the job for which they are evaluated or required. To justify such relationships, more than a superficial resemblance between the content of the personal history variables and the content of the job is required. For example, things like course titles and job titles may not give an adequate indication of the content of the course or the job or the level of proficiency an applicant has developed in some critical area. The critical consideration is the similarity between the behaviors, products, knowledges, skills, or abilities demonstrated in the personal history variable and the behaviors, products, knowledges, skills, or abilities required on the job, whether or not there is a close resemblance between the given personal history variable as a whole and the job as a whole.

D. Job Content Domains Should be Defined in Terms of What an Employee Needs to Do or Know Without Training or Experience on the Job. It is important to delineate the knowledge, skills, and abilities an employee is expected to have before placement on the job, and to define the selection domain in those terms. There is a fine line between what employees bring to the job and what they are taught in training or on the job. In many instances, those who bring more learning to the job require shorter or different training than others. The researcher should seek the appropriate balance between selection and training and define the content domain for the procedure in accordance with this balance (Goldstein, 1980). For example, the fact that an employee is taught to interpret company technical manuals does not mean that the job applicant should not be evaluated for basic reading skills. Although, a test of basic reading skills would not be defensible on content grounds alone; a sample of the material in the technical manual may be considered content valid. Those job candidates who are selected should have the knowledge and skills required to learn to do the job as well as those required to perform the job.

E. A Job Content Domain May be Restricted to Critical or Frequent Activities or to Prerequisite Knowledge, Skills, or Abilities. There is no virtue in measuring ability to handle trivial aspects of work or trivial knowledge. On the other hand, a single activity may make up so much of the job that the job may be considered a single domain for measurement purposes. For example, truck drivers must be able to drive a truck. The fact that they may perform other functions is irrelevant to developing and using a measure of driving skill or ability. A selection procedure developed on the basis of job content domain sampling should not include content irrelevant to the domain

sampled. This condition is almost impossible to meet since virtually any content-oriented procedure includes some elements that are not part of the job domain; for example, standardization of the selection procedure, use of multiple choice formats, or requirements to read. Irrelevant components should be minimized.

F. Sampling of a Job Content Domain Should Ensure that the Measure Includes the Major Elements of the Defined Domain. The process of constructing or choosing the selection procedure requires sampling the job content domain. The sampling done at this stage should have some rationale; that is, there should be rules for the inclusion of critical elements. The acceptability of the claim of content validity of the selection procedure rests on the extent to which elements of the selection procedure match elements of the job content domain. The acceptability of the match between selection procedure and job content domain is a matter of professional judgment.

G. A Test Developed on the Basis of Content Sampling Should Have Appropriate Measurement Properties.
1. Where feasible, the selection procedure should be pretested. The report should provide estimates of the means, variances, and intercorrelations of parts (i.e., items or sets of items) of the procedure.
   a. Parts that do not contribute to the total variance should ordinarily be eliminated. They may be replaced with tests that do differentiate, but care should be taken so that the content domain of the replacement test is the same as that of the test eliminated. This particular principle may not hold when the researcher is interested in the mastery of a particular critical skill or knowledge.
   b. Items should be chosen on the basis of item statistics to achieve desirable overall test characteristics but always with consideration of the degree to which the items represent appropriate content.
   c. Extreme intercorrelation of parts amounting to redundancy of measurement should be avoided. Inclusion of items in a content valid test should be based primarily on their relevance to some content domain, and secondarily, on their intercorrelations. Well constructed parts that do not correlate with other parts or a total score should not necessarily be eliminated. If the lack of correlation among parts of the selection procedure reflects the lack of correlation of parts of the content domain, it is appropriate to include the uncorrelated parts in the selection procedure.
2. Reliability is a matter of concern in all measurement, but it is a particular concern for work samples. Equipment may wear or
function variably; scoring variations may occur; a desire to minimize testing time may result in taking a sample of performance too small to ensure reliable results; practice and fatigue effects may also be a problem.

3. When the primary objective of the selection procedure is the classification of individuals into pass-fail groups, the emphasis should be on the reliability of the assignment of individuals to these categories.

Pass-fail decisions are the objective of criterion referenced tests which have been used widely in educational contexts and less frequently in personnel selection (Buck, 1975). Use of criterion-referenced measures raises additional measurement issues (Hambleton, Swaminathan, Algina & Coulson, 1978). A set of guidelines for criterion referenced tests and test manuals is contained in Hambleton and Eignor (1978) and a critique of standard-setting has been provided by Glass (1978).

4. Scoring keys for content-oriented tests should be checked for accuracy just as in any other test. Answers that are keyed as correct should be correct under all reasonably expected job-relevant circumstances. Those choices or potential answers keyed as incorrect should be incorrect under all reasonably expected job-relevant circumstances.

5. Interpretation of content-oriented selection procedures depends on the measurement properties of the given procedure. If a selection instrument measures a substantial and important part of the job reliably, and provides adequate discrimination in the score ranges involved, persons may be ranked on the basis of its results. However, if an instrument is constructed more in the manner of a training mastery test, in which the examinee is expected to get all or nearly all of the items correct, a critical score may be appropriate. A critical score is also appropriate in situations such as those in which greater speed or accuracy cannot be reflected in production because of equipment or process limitations. In that case, the limiting conditions should be considered in the design of the selection procedure.

H. Persons Used in Any Aspect of the Development or Choice of Selection Procedures to be Validated on the Basis of Content Sampling Should Clearly be Qualified. Panels of job experts (i.e., people with thorough knowledge of the job(s) ) may be used in defining domains, in writing test items, in developing simulation exercises, and in evaluating items or total procedures. Any individuals involved in the procedure construction or choice process should be thoroughly trained in those aspects of measurement necessary for their roles.

---

Construct-Oriented Strategies

Psychological characteristics such as reasoning ability, spatial visualization, sociability, or introversion are called constructs because they are theoretical constructions about the nature of human behavior. The early concept of a construct (Cronbach & Meehl, 1955) implied a sophisticated, highly-developed, and continually evolving theory; current usage generally looks upon the idea of a construct less elegantly. Ghiselli, Campbell, and Zedeck (1981) identified a construct as "a characteristic of individuals that many people have deemed important—important enough to define and measure" (Ghiselli et al., 1981, p. 10). If constructs are deemed important for personnel selection, they are important enough to measure validly. Knowing whether a construct is measured validly requires, if not a theory, at least some fairly well articulated ideas about what is being measured, what a measure of the construct should reasonably be expected to be related to and, perhaps more importantly, what it should not be related to, (i.e., what might be a contaminating source of variance). This view of constructs and construct validity implies two aspects of a construct-related strategy for developing evidence to judge the job relatedness of a selection procedure. The first is evidence that the construct is indeed important for job performance—that is, evidence must be grounded in a thorough knowledge of the job. Ordinarily, a job analysis can provide a part of the basis for identifying and defining psychological constructs which are important to job performance. Clarity of the articulation of the meaning and the nature of the construct, and well-informed expert judgment that a logical relationship exists between the nature of the construct and identifiable demands of the job is essential.

The second is evidence that the instrument used as a selection procedure is a valid measure of the construct and not of other constructs. A construct validation strategy for evaluating personnel selection procedures requires attention to the development of both kinds of evidence.

The standards of construct validity in personnel selection have not been as clearly delineated as they have been for content and criterion-related strategies. At a minimum they include the following principles, the first of which is quoted from the Standards:

1. "The construct for a particular test should be embedded in a conceptual framework, no matter how imperfect that framework may be. The conceptual framework specifies the meaning of the construct, distinguishes it from other constructs, and indicates how measures of the construct should relate to other variables" (AERA et al., 1985, pp. 9-10).

2. There should be evidence that the test adequately measures the construct under consideration. First, the measure should be internally consistent. Second, the rationale for identifying and measuring the
construct should be supported by patterns of correlations or experimental evidence which show that the tests measure the construct under consideration and not other constructs.

3. The compilation of construct-related evidence for test validity begins with test development and continues until the pattern of empirical relationships between test scores and other variables clearly indicates the meaning of test scores.

Construction or selection of the assessment procedure is based on a thorough understanding of the work and of what the worker is expected to accomplish. With such understanding, the researcher generates hypotheses about the characteristics needed to do the job well. These hypotheses should be informed not only by the understanding of the job and its products but also by evidence in the research literature regarding the relationship between similar kinds of work or important aspects of the work and the construct or related constructs. The researcher should articulate a clear conceptual rationale for the linkage between the construct and the work to be done.

4. The practice of asking subject matter experts to list the constructs that they believe are required for successful performance does not, without further evidence, provide a basis for a claim of construct validity. These judgments must also be based on knowledge of worker characteristics required for adequate job performance and how they are measured by the proposed selection procedure.

Validity Generalization

Classic psychometric teaching has held that validity is specific to the research study and that inability to generalize is one of the most serious shortcomings of selection psychology. Current research has shown that the differential effects of numerous variables are not so great as heretofore assumed; much of the difference in observed outcomes of validation research can be attributed to statistical artifacts (Callender & Osburn, 1980; 1981; Hunter & Hunter, 1984; Schmidt, Hunter, & Pearlman, 1981). It now seems well established from both validity generalization studies and cooperative validation efforts that validities generalize far more than once supposed.

Differential prediction, in the sense of efforts to find different selection procedures for predicting different aspects of performance in a job or for predicting performance in different jobs, has proven to be difficult. This, too, is evidence that validities are more generalizable than has usually been believed.

Validity generalization is a demonstration that a selection procedure or kind of selection procedure permits valid inferences about job behavior or job performance across given jobs or groups of jobs in different settings.

Systematic meta-analyses have been used to integrate the cumulative findings from a number of validity studies to determine the best estimates of the validity of the procedure for the kinds of jobs or groups of jobs and settings included in the studies (Schmidt, Hunter, Pearlman, & Shane, 1979; Callender & Osburn, 1980; Raju & Burke, 1983; Linn & Dunbar, 1985; Schmidt, Pearlman, Hunter, & Hirsch, 1985; Sackett, Schmitt, Tenopyr, Kehoe, & Zedeck, 1985). Most of these efforts have involved the validity of cognitive ability tests; meta-analyses regarding other test-job performance relationships are also now being performed.

To establish validity generalization for a particular selection procedure or kind of selection procedure, there must be evidence that demonstrates either (a) that situational differences have little or no effect on operational (or true) validities or (b) that true validities of measures of the construct or of the predictor type have acceptably low probabilities of falling below specified values in new settings. Evidence that situational differences have little effect may take the form of an accumulation of study results indicating approximately the same true validity for a given selection procedure regardless of situational factors. In evaluating such evidence, the researcher should consider the likely probabilities of both Type I and Type II decision errors (Sackett, Harris, & Orr, 1986). Evidence that true validities fall above a specified lower limit may take the form of research results showing that the lower bound of the confidence interval of a distribution of true validities is positive.

To the extent that validity generalization evidence is available, researchers may rely on it to support the use of selection instruments. Researchers and employers are encouraged to conduct cooperative studies when adequate data for validity generalization are not available.

Because of the complexity of methods and findings in this area, professional judgment in interpreting and applying results is important. Researchers should consult the relevant research literature to ensure that the procedures to be used have been professionally and soundly developed, that the appropriate procedures for estimating validity on the basis of cumulative evidence have been followed, that the conditions for the application of validity generalization results have been met, and that the application of validity generalization conclusions are appropriate for the jobs or groups of jobs and settings studied. The rules by which researchers categorized the jobs studied, the tests, the criteria, and other study characteristics should be fully reported as well as the reliability of the coding of these data. Citations to or copies of the reports that contribute to the meta-analytic results should be available. The quality of the individual research studies and their impact, if any, on the validity generalization conclusions and their use should also be informed by good professional judgment.
The cumulative validity evidence may indicate generalizability of validity for a selection procedure only for particular kinds of jobs or job families. In such a case, reliance on validity generalization results for jobs in new settings or organizations should meet two conditions:

1. The selection procedure to be used is a measure of the trait, ability, or construct studied, or is a representative example of the type of selection procedure included in the validity generalization study. Inferences about job performance or training performance made on the basis of test scores should be supported by the evidence produced from the validity generalization study.

2. The job in the new setting is similar to the job, or a member of the same job family, included in the validity generalization study. A wide variety of methods is available to examine inter-job similarity or job family membership for validity generalization purposes. If sufficient information is available on the new job families to accurately assign it to a relevant occupational category, classification can be made without a formal job analysis. The cumulative evidence from validity generalization studies is more reliable and accurate than the evidence from any single study. However, sole reliance upon available cumulative evidence may not be sufficient to meet such employer needs as the selective placement of workers and the optimal combination of procedures. Additional studies may be required to meet such information needs. If such studies are not feasible in one's organization, researchers and employers are encouraged to engage in cooperative studies to accumulate the evidence needed to ensure reliable results.

Meta-analytic techniques are still evolving and researchers should be aware of continuing research that may provide further refinement of the techniques as well as a broader range of test-performance relationships to which the technique has been applied. Researchers should be aware of continued work and critiques regarding meta-analysis methods (Bullock & Svyantek, 1985; Burke, 1984; Light & Pillemer, 1984; Orwin & Cordray, 1985), as well as continued application of meta-analytic techniques to investigate the validity generalizability of selection procedures other than cognitive ability tests.

**Operational Use**

Validation is the investigatory phase in the developmental or choice of selection procedures. Whatever the outcome of validation research, the researcher should prepare a report which documents critical aspects of the research and the findings of that research. The researcher should provide detailed information on how to use the procedure, usually in a procedures manual, directions on operational use are almost always necessary. Details on the research and development of the procedure are useful in the event of any subsequent need to modify the technique, update the research supporting the selection procedure, replicate a previous study, or defend the appropriateness of the selection procedure.

A. **Research Reports and Procedures Manuals.** Whether or not a separate research document and procedures manual are written depends on the needs of the organization. In any case, the following information regarding the procedure should be available when needed:

1. **Identifying information.** The author and date of the study or other information that would permit retrieval of information at a later date should be provided.

2. **Problem.** The deficiencies in previous selection procedures which the new procedure is designed to rectify or improvements in employment which the new procedures are expected to accomplish should be stated.

3. **Job Description.** The report should include a statement of the results of the job analysis.

4. **Criterion Measures.** A description of the measures, the rationale behind their use, the data collection procedures, and a discussion of the relevance, reliability, and freedom from contamination of the resulting criteria should be described in detail. [Not relevant for content-oriented strategies.]

5. **Research Sample.** Characteristics of the sample relevant to an interpretation of the results, should be described. The description should include a definition of the population which the sample is designed to represent, sampling biases which may detract from the representativeness of the sample, and the significance of any deviations from representativeness for the interpretation of the results. Data regarding restriction in the range of individual differences on predictors or criteria are especially important.

6. **Selection Procedures.** Names and editions of published procedures should be provided as well as descriptions, and, if possible, sample items or copies of tests that are not generally available to professionals.

7. **Relationship to job duties.** (Particularly in reports of content validation strategies). The report should provide a description of the means by which it was determined that the selection procedure is representative of a domain of performance on the job, or of the knowledge or skills that are prerequisites to job performance.

8. **Identification and results of related studies.** Complete references to published studies, and copies of unpublished studies which are relied on should be available.

9. **Results.** All summary statistics that bear on the conclusions drawn by the researcher and the recommendations should be in-
14. Tables should present complete data, not just significant or positive results. Both uncorrected and corrected values should be presented when corrections are made for artifacts such as restriction of range or unreliability of the criterion. The sample size, and means, standard deviations and intercorrelations of variables measured and other information useful to the interpretation of the results should be presented and clearly labeled.

10. **Recommendations.** The recommendations and the data and rationale supporting the recommendations, including, but not limited to rank ordering, banding of scores, cutoff scores, and the means of combining information in making personnel decisions should be provided.

11. **Informational Material.** Information developed for users or examinees should be accurate and complete for its purposes, and should not be misleading. The reading level of the material prepared for the applicants should match their reading ability. Memoranda and management records should be worded to communicate as clearly and accurately as possible the information that readers need to know to carry out their responsibilities competently and faithfully. Care must be taken in preparing such documents to avoid giving the reader an impression that an assessment program is more useful than it really is.

12. **Cautions Regarding Interpretations.** Research reports or procedures manuals should help readers make correct interpretations of data and should warn them against common misuses of information.

13. **Qualifications or Training of Administrator.** Any special qualifications required to administer a selection procedure or to interpret the scores or other measurements should be clearly stated in the research report and/or procedures manual.

14. **Instructions and Training for Administrators.** The procedures manual should specify the procedures to be followed and emphasize the necessity for standardized administration, scoring, and interpretation. These instructions should be so clear that all persons concerned will know what they are to do. Everyone involved should understand that failure to follow standardized procedures may render the research results irrelevant to some degree. The researcher must be both insistent and persuasive to get people to understand both the nature of and the need for standardized administration of tests or other procedures. Periodic seminars run by the researchers or by other appropriately trained professionals may be needed to reinforce the written instructions. Observational checks or other quality control mechanisms should be built into the system. There may be situations where research is based on data from operational studies where unstandardized procedures have been used with no serious impairment of validity. In such situations, the standardization is relatively unimportant. However, the acceptability of unstandardized procedures should not be assumed without investigation.

15. **Scoring Procedures.** Instructions for scoring or scaling should be presented in the procedures manual in as much detail as needed to reduce clerical errors in scoring and to increase the reliability of any judgments required. When keys must be kept confidential, they should be made available only to persons who use them to score or scale the responses.

16. **Normative Information.** As stated above, a research report should contain clear and complete descriptions of the samples used in the research. Normative reporting should include measures of central tendency and variability and should clearly establish the nature of the normative data given (e.g., centiles, standard scores). Ordinarily, norm tables are less useful than expectancy tables for employment decisions. Expectancy tables indicate the proportion of a specific sample of candidates who reach a specified level of success. Norm tables may be useful in determining the effects of a cutting score.

17. **Transformation of Raw Scores.** Any derived scale used for reporting scores should be described carefully in the research report or procedures manual. Whether using derived scores (such as those described in general textbooks on measurement) or locally produced scales (such as "qualified," "marginal," or "unqualified"), the researcher should make clear their logical and psychometric foundations.

18. **Review.** Research should be brought up to date and reported, and procedures manuals should be revised as needed. Any changes in the research data or use of the procedures that would make any statement or instruction in such documents incorrect or misleading should result in revision.

In summary, whenever a selection procedure is made available for use in personnel decisions, one or more documents should be prepared to describe validation research and the standard procedures to be followed in using the results of that research. Reports of validation research should include enough detail to enable a researcher competent in personnel selection to know what was done, to draw independent conclusions in evaluating the work, and to replicate the study. The reports must not be misleading. Research findings which might qualify the conclusions or the generalizability of results should be reported. Procedures manuals should...
provide adequate information to the test user to insure continued appropriate administration and use of the selection procedure.

B. Use of Research Results. The researcher should consider the use of the research and the reported findings by the employer. The following principles apply:

1. It is the responsibility of the researcher to provide useful methods for interpreting the test scores to the user(s). Although the management of the organization usually retains the final decision on the use of a specific selection procedure, it is the responsibility of the researcher to recommend how the procedure is to be used. The recommended use should be consistent with the procedures used in validation and the results of the study.

2. Computer Based Test Interpretations (CBTI) should be validated using the principles outlined in this document. Whether a CBTI provides a decision for the user or a decision-maker uses the information along with other data to make the decision, it is the final decision that should be validated. Although there may be legitimate need to withhold computer scoring algorithms because of proprietary concerns, the employer is still expected to validate the selection procedure in ways consistent with these Principles. The American Psychological Association (1986) has developed guidelines that may be informative in the use of CBTI.

3. Cutoff or other critical scores may be set as high or as low as the purposes of the organization require, if they are based on valid predictors. This implies that (a) the purposes of selection are clear and (b) they are acceptable in the social and legal context in which the employing organization functions. In usual circumstances, the relationship between a predictor and a criterion may be assumed to be linear. Consequently, selecting from the top scorers on down is almost always the most beneficial procedure from the standpoint of an organization if there is an appropriate amount of variance in the predictor. Selection techniques developed by content-oriented procedures and discriminating adequately within the range of interest can be assumed to have a linear relationship to job behavior. Consequently, ranking on the basis of such scores is appropriate. It is not necessary to add any assumptions about underlying traits in order to rank the applicants. In some circumstances, such as those where a production line limits the speed at which a worker can produce, a critical score may be in order.

Judgment is necessary in setting any critical or cutoff score. A fully defensible empirical basis for setting a critical score is seldom, if ever, available. The only justification that can be demanded is that critical scores be determined on the basis of a rationale which may include such factors as estimated cost-benefit ratio, number of openings and selection ratio, success ratio, social policies of the organization, or judgments as to required knowledge, skill, or ability on the job. If critical scores are used as a basis for rejecting applicants, their rationale or justification should be made known to users.

4. Employers should provide opportunities for reconsidering candidates whenever reconsideration is technically feasible. Under at least some circumstances, employers should allow candidates to reapply. There are several possible reasons for questioning the appropriateness of prior assessment for an applicant. One such consideration might be length of time since prior assessment. Where there is evidence of relevant new learning, retesting or reevaluating is usually a desirable practice.

5. The use of a predictor which is not supported by cumulative research literature should be accompanied by systematic procedures for developing additional data for continued research. The operational use of a valid predictor may result in such severe restriction of range that its validity cannot be demonstrated in subsequent research (Peterson & Wallace, 1965). There is no well-established general procedure for determining the validity of instruments which have been used to select persons who become the research sample. Researchers should develop strategies to study the possibility that validity decreases over time.

6. All persons within the organizations who have responsibilities related to the use of employment tests and related predictors should be qualified through appropriate training to carry out their responsibilities. The person in charge of a selection program should understand the principles of measurement, the limitations on the interpretations of test scores and other aspects of the literature relevant to the use of selection procedures and employment problems. Persons in the organization who have responsibilities related to the selection program should have the training necessary to carry out those responsibilities competently.

7. Researchers should seek to avoid bias and even the appearance of bias in choosing, administering, and interpreting selection procedures. At the very least, the use of a selection procedure should create an environment that is responsive to the feelings of all candidates, insuring the dignity of all persons.

8. There should be a periodic audit of the use of selection procedures. Departures from established procedures often occur over time and new findings in psychological or psychometric theory may be relevant to one or more of the assessment procedures in
10. Not all findings of differential prediction should lead to different uses of a predictor for different groups. For example, if the study is based upon an extremely large sample, a finding of small but statistically significant differential prediction may have little practical impact. Occasionally, data indicating differential prediction may result from statistical artifacts or may suggest courses of action inconsistent with societal goals. In the latter case, the reasonable course of action would probably be to recommend uniform operational use of the prediction for different groups, or perhaps, to conduct further research.

11. Confidential test materials should be adequately safeguarded. Decision makers should beware of basing decisions on scores obtained from compromised selection procedures. This principle is difficult to apply to non-test predictors such as judgments reached in an employment interview. Nevertheless, the importance of security as a means for preserving standardization and validity should be considered. Selection procedure scores should be released only to those who have a need to know, and who are qualified to interpret them. The results of reference checks, for example, should be held confidential. Precautions should be taken to avoid the disclosure of information which may harm the applicant or damage the program.

12. Procedures should be controlled so as to safeguard the validity of the selection procedures. Any prior information given to candidates about the selection procedures should be uniform for all persons. Particular care should be taken so that some individuals do not have advantages, such as coaching, that were not present during the validation effort. If these practices occur, their impact on validity should be evaluated. Finally, public disclosure of the content of most selection procedures should be recognized as a serious threat to their validity, reliability, and subsequent development. Work samples and physical ability tests may be exceptions to this general statement.

13. Scores may lead to invalid inferences because of unusual features of the situation (e.g., uncommon distractions), exceptional characteristics of the individual (e.g., physical handicap) or the passage of time (e.g., relevant new learning since evaluation occurred). When unusual circumstances exist, the decision rules should be reevaluated in the light of the additional information. The principle is that some degree of judgment should be retained in the interpretation of scores obtained in circumstances markedly different from those in the validation research.

14. Raw scores should be kept, since data reported in derived scales may be inappropriate for further research.

15. Information from a selection procedure should only be used for decisions for which it was intended. A particular problem is the use of entry-level test information or old information of other kinds retained in files. For example, although selection test data may have some validity in determining later retention decisions, better measures such as performance ratings may be available. However, appropriate data should be retained separately for future research.

16. Results should be reported in language likely to be interpreted correctly by persons who receive them. Scores should not be reported to candidates or to managerial personnel unless they are explained carefully to make certain that interpretations are correct. Scores should not be reported to persons who may be asked later to provide criterion ratings for validation. If such studies are contemplated, provision should be made from the outset to make sure that the raters have no access to the scores, and as far as possible, are unable to guess them.

17. Derived scores, such as I.Q.'s or grade-equivalent scores, which are often terms used in the educational context, are to be avoided. These terms are easily misinterpreted and are not likely to be directly meaningful for making employment decisions. Even where they had legitimate psychometric significance historically, they have been so encrusted with spurious meaning that they lend themselves to misinterpretation.

18. Selection procedures should be administered only where there is a need to know the scores of the subjects. Casual administration of selection procedures to supervisors and others who have no real need to take them can result in breaches of security and can cause damage. This principle does not preclude administration for research purposes under appropriately controlled conditions.

Summary

These Principles are meant to specify good practice in the choice, development, evaluation, and use of personnel selection procedures. They present currently accepted practice in personnel selection and, as such, are not mandates, nor do they specify minimum standards.
A review of the changes which have taken place in the Principles since the first edition in 1975 should assure the reader that the field of personnel selection is far from static. Certain issues for which the research base is still evolving have not been treated in complete detail in the Principles. Examples include job analysis, methods of using job analysis in the development of measuring instruments, methods of determining utility, information on psychological constructs, and validity generalization. The Principles focus primarily on the establishment of the validity of a single selection procedure. In most practical situations, multiple procedures are used and must be combined to yield a decision. The fundamental issue is not how the data are combined, but the validity of the decision.

Furthermore, many relatively new developments in measurement are not covered extensively (e.g., generalizability theory for analysis of instrument reliability, computer adaptive testing, and criterion-referenced testing). Some of these developments as well as others are likely to prove useful for selection. Hence, practitioners should be aware of the possible value of applying such techniques in personnel selection.

Finally, although the Principles are not designed to generate debate, the Society welcomes useful commentary that may aid in the development of future editions.

Glossary

**Behavior**: observable aspects of a person's activities.

**Bias**: a sample of measures is biased relative to population of measures if some elements from the population are disproportionally represented.

**Claudy formula**: an empirical formula to correct for shrinkage in cross-validation, including the biased, nonrandom selection of a smaller number of predictors from a larger number given.

**Concurrent validity**: a demonstrated relationship between job performance and scores on selection instruments obtained at approximately the same time.

**Confidence interval**: the bounds on a measurement defined by a certain probability that the parameter of interest will be included.

**Configural scoring**: the assignment of weights to paired variables so that the inference to be drawn from one predictor score depends upon the level of the second predictor score.

**Construct**: a trait of individuals inferred from empirical evidence or theory.

**Construct-oriented strategy**: an attempt to demonstrate a relationship between underlying traits or hypothetical constructs and job related behavior.

**Contamination**: a measure is contaminated if irrelevant sources of systematic variance account for substantial portions of the total variance.

**Content domain**: a body of knowledge and/or set of tasks or other behaviors defined so that given knowledge or behaviors may be classified as included or excluded.

**Content-oriented strategy**: an attempt to establish that test performance is a representative sample of job performance or job-required knowledge.

**Correlation**: the degree to which two sets of measurements vary together.

**Criterion**: a measure of job performance or job behavior, such as productivity, accident rate, absenteeism, tenure, reject rate, training score, and supervisory or co-worker ratings.

**Criterion-related strategy**: an attempt to demonstrate a statistical relationship between scores on a predictor and scores on a criterion measure.

**Critical score**: a specified point in a distribution of scores at or above which candidates are considered successful; differs from a cutoff score in that a critical score is the same for all applicant groups.

**Cross validation**: the application of a scoring system or set of weights empirically derived in one sample to a different sample from the same population to investigate the stability of relationships based on the original weights.

**Cutoff score**: a point in a predictor distribution below which candidates are rejected; differs from critical score in that a cutoff score may be dependent on the number of openings and the number of applicants.